Stony Brook University

The making of a generalist: mining the transcriptome to characterize host-use evolution in the aphid *Uroleucon ambrosiae*.

A Dissertation Presented

by

Aman S. Gill

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Ecology & Evolution

Stony Brook University

August 2014

Stony Brook University
The Graduate School

Aman S. Gill

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Douglas J. Futuyma – Dissertation co-advisor
Distinguished Professor, Department of Ecology and Evolution

Joshua Rest - Dissertation co-advisor
Assistant Professor, Department of Ecology and Evolution

Walter F. Eanes, Chairperson of Defense
Professor, Department of Ecology and Evolution

Daniel J. Funk
Associate Professor, Department of Biological Sciences, Vanderbilt University.

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

The making of a generalist: mining the transcriptome to characterize host-use evolution in the aphid *Uroleucon ambrosiae*.

by

Aman S. Gill

Doctor of Philosophy

in

Ecology & Evolution

Stony Brook University

2014

The evolution of host use—the preference and ability to utilize plant species—has long occupied a central role in theories on how and why insects are among the most diverse lineages on earth. Yet the genomic basis of host-range evolution remains poorly understood, partly because most studies concern species-level comparisons, where reproductive isolation is already complete, obscuring the evolutionary changes that may have initiated divergence. To advance our understanding of host-range evolution at the instraspecific, genomic level, I generate and investigate genomic data for populations of the aphid *Uroleucon ambrosiae* (*Ua*) that diverge in host use traits—eastern North American populations specialize on giant ragweed, while those in the arid southwest exploit a broad range of host plant genera. Three sets of analyses were used to clarify the host-use history of *U. ambrosiae* and characterize genomic divergence among its generalist and specialist populations. First, multi-locus sequence data and host records are used to resolve the phylogeny of the genus, and establish that *Ua* is part of an adaptive radiation of species feeding on Asteraceae tribes that began 0.5-2 million years ago. Second, the population-level transcriptome is assembled and annotated, and compared to the pea aphid genome. These results indicate that lineage-specific gene expansion has likely been an important factor as *Uroleucon* species diversified on host plants in the family Asteraceae. Finally, expressed sequences are compared among multiple generalist and specialist *Ua* populations. Genes that possibly contribute to functional divergence of host use traits are identified using a combination of outlier analyses based on pair-wise allele frequency divergence (D), the fixation index (F_{ST}), and the ratio of non-synonymous to synonymous substitutions (K_a/K_s). These results identify several candidate genes for divergent host use, including constituents of salivary secretions as well as several genes involved in digestive metabolic processes.

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

Acknowledgments

This work could not have been completed without my dissertation committee. Dr. Douglas Futuyma provided inspiration for working on questions regarding ecological specialization and plant-insect interactions. Dr. Joshua Rest provided guidance, technical expertise, and support for laboratory work throughout the process. Funding for this work was provided by funds from the Rest and Futuyma laboratories, as well as several small grants administered by the Department of Ecology & Evolution, the Graduate Student Organization, and the Graduate Student Employees Union, all of Stony Brook University. Dr. Daniel Funk generously made available background material and field information on my study species, which he had previously studied during his postdoctoral years. Dr. Walt Eanes provided criticial insight at key turns of the project. I also benefitted from the help of fellow graduate students. Niamh O'hara, Stephen Sabatino, Megan Flenniken, Adam Ehmer, and Patty Oikawa all supported my work in their own way. I also thank my family. My dad, Bikram Gill, has been an inspiration through example as a worldwide leader in wheat research. My mom, Billo Gill, has always been behind me, and even served as my field assistant for one Arizona collecting trip. My siblings Jason, Naseeb, and Ragini, and my sistern-in-law Annita have been interminably supportive. Artemis Gill's arrival provided a final push. I'm grateful for the friendships that have sustained me. Vinson Doyle, Oren Tzfadia, Chris Brunson, Denise Fillion, Laura Davis, Karen Weingarten, Svetlana Kitto, Maryan Soliman, Ben Newman, Idan Ben-Arieh, and Stephanie Higgins are all people I value in my life and my work. My number one, Kristin Overton, has been my biggest supporter throughout.

Chapter 1

Overview

The evolution of host use among insect herbivores has long occupied a central role in theories of how and why insects have become among the most diverse lineages on the planet (Ehrlich & Raven, 1964; Futuyma & Agrawal, 2009; Winkler & Mitter, 2008). Macroevolutionary patterns of host-use indicate that most insect herbivores specialize on closely related host plant groups (Schoonhoven *et al.,* 2005), although transitions are common through time between specialization and more generalized host-breadth involving relatively indiscriminate use of unrelated hosts (Nosil, 2002). Host-use transitions may encourage diversification, as related species commonly experience host-associated ecological divergence (Shafer & Wolf, 2013; Winkler & Mitter, 2008), which is positively correlated with reproductive isolation across disparate lineages (Funk *et al*., 2006). A major obstacle in resolving a possible mechanistic relationship between host-use evolution and diversification is our lack of detailed understanding of the genomic basis of host-use divergence at the micro-evolutionary level, *i.e.* among populations of a single species. In this dissertation I investigate the aphid *Uroleucon ambrosiae* (*Ua*) as a system for the study of intra-specific ecological divergence.

Chapter 2 provides phylogenetic context for the study of *Ua* populations. Multi-locus gene sets are used to resolve phylogenetic relationships of *Uroleucon* species along with other members of their tribe, the Macrosiphini. Host-use characteristics are mapped to the resulting phylogeny in order to reconstruct ancestral host-use traits. These analyses support two major findings. First, the *Uroleucon* lineage switched to feed on its primary host plant family, Asteraceae, after it diverged from its most recent common ancestor with the model species *Acyrthosiphon pisum*, the pea aphid. Second, *Ua* is part of a North American adaptive radiation that began 0.5-2 million years ago. Species in this radiation variously specialize on hosts in several different Asteraceae tribes, all of which are utilized by generalist populations of *Ua*. These results suggest that generalist *Ua* populations may harbor ancestral genetic variation to utilize hosts that are not used by specialist populations.

This study presents the first genomic resources for *Ua.* Chapter 3 describes sequencing, *de novo* assembly and annotation of the *Ua* transcriptome, based on field-collected samples from

across North America. Comparative genomic analysis of the *Ua* transcriptome with the pea aphid genome allows identification of over four hundred host-use associated genes spanning multiple functional categories, including chemosensation and salivary secretion constituents. The *Ua* phylome is also presented, providing a catalog of protein-based phylogenetic trees that resolve relationships between *Ua* coding sequences and homologous proteins from other insect species. These data are used to identify gene families that have undergone lineage-specific gene expansion in *Ua* or its ancestors. These include several gene families of relevance to host-use, including salivary gland genes and inhibitor proteins that target plant proteinase inhibitors, which can disable aphid digestive enzymes. Chapter 2 also reports on experimental treatments designed to identify genes differentially expressed in response to different host plants. Several clusters of co-expressed genes are identified, including purine metabolism genes that may be involved in integrated metabolic responses of *Ua* and its bacterial endosymbiont *Buchnera aphidicola*.

Chapter 4 focuses on intra-specific analysis of specialist and generalist *Ua* populations, drawing on the genomic resources described in Chapter 3 and the phylogenetic context provided in Chapter 2. The primary objective pursued in Chapter 4 is to use patterns of variation of single nucleotide polymorphisms (SNPs) to identify candidate loci for host-use divergence in generalist versus specialist populations. Analysis of SNPs—including calculation of allele frequency differences among populations (D) and the fixation index (F_{ST})—points to outlier loci that are differentiated in pair-wise comparisons of generalist and specialist populations, but not generalist-generalist or specialist-specialist comparisons. A distinct method to identify candidate loci, based on the ratio of non-synonymous to synonymous substitutions (K_a/K_s), pinpoints loci with elevated rates of protein-changing (*i.e.* non-synonymous) substitutions. The possible functional role of candidate genes is discussed. Notably, these include sucrase and other digestive enzymes targeting components of plant phloem, proteins likely to be involved in the aphid-*Buchnera* symbiosis, and, as in chapter 2, proteins targeting plant proteinase inhibitors.

This project illustrates a methodology to develop population-level genomic resources in a non-model system to identify candidate genes for functional ecological traits. The findings point to several avenues for follow-up research, including validation of candidate loci through experimental manipulation and targeted re-sequencing based on genomic DNA samples representing a more dense sampling of *Ua* populations from across their range.

CHAPTER 1 REFERENCES

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and plants: a study in coevolution. Evolution, 586–608.

Funk, D. J., Nosil, P., & Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. Proceedings of the National Academy of Sciences, 103(9), 3209–3213.

Futuyma, D. J., & Agrawal, A. (2009). Macroevolution and the biological diversity of plants and herbivores. Proceedings of the National Academy of Sciences, 106(43), 18054-18061.

Nosil, P. (2002). Transition rates between specialization and generalization in phytophagous insects. Evolution, 56(8), 1701–1706.

Shafer, A. B. A., & Wolf, J. B. W. (2013). Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. Ecology Letters, 16(7), 940–950.

Schoonhoven, L. M., van Loon, J. J. A., & Dicke, M. (2005). Insect-Plant Biology. Oxford University Press.

Winkler, I., & Mitter, C. (2008). The phylogenetic dimension of insect-plant interactions: a review of recent evidence, 240–263. In: K.J. Tilmon (Ed.), Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects. University of California Press.

Chapter 2

Biogeography and evolution of host associations in *Uroleucon* aphids

INTRODUCTION

The history of aphid diversification is replete with shifts in associations between insects and host plant lineages, not unlike many other insect groups (Blackman & Eastop, 2008; Peccoud *et al.*, 2010). The history of aphid host use reflects several patterns found in most herbivorous insect lineages, the result of a long-standing emphasis on the macroevolution of insects and plants (Futuyma & Agrawal, 2009; Winkler & Mitter, 2008). Herbivorous insect clades are consistently more diverse than their non-phytophagous sister groups (Mitter *et al.,* 1988). In many groups of insects, host use is phylogenetically constrained, such that insects are more likely to feed on closely related host plant species than not, and most related insects utilize related hosts (Blackman & Eastop, 2008; Peccoud *et al.*, 2009). Yet shifts to dissimilar hosts have occurred at various times in the history of most lineages (Futuyma & Mitter, 1996; Mitter *et al.*, 1991; Winkler & Mitter, 2008). Most species specialize on plants in a single family and not uncommonly a single genus. Despite its preponderance, the evolution of host specialization is not a one-way process; transitions from specialism to generalism are about as frequent as the reverse (Futuyma & Moreno, 1988; Kelley & Farrell, 1998; Nosil, 2002; Winkler & Mitter, 2008).

These patterns together suggest that co-evolutionary interactions of insects and plants—and host-use evolution in particular—play a functional role in species-level diversification*, i.e*. the evolution of reproductive barriers among formerly conspecific populations (Mitter *et al.*, 1988). Most research to date has focused on the host-use characteristics of species, or "host races," that have already become reproductively isolated. To further advance knowledge of insect-plant interactions at the mechanistic or microevolutionary level, we build on previous studies on the aphid genus *Uroleucon* and its tribe, the Macrosiphini (Moran *et al.*, 1999; Moran, 1984; Robinson, 1985; Robinson, 1986). In particular, we focus on the phylogenetic and host-use history of Nearctic members of the sub-genus *Uroleucon*, with a focus on its best-studied representative, *U. ambrosiae (Ua)*.

Populations of *Ua* diverge in host-use traits, ranging from specialist to generalist roughly along an east-west geographic gradient (Bernays & Funk, 2000; Bernays *et al.*, 2000; Funk & Bernays, 2001). This feature makes it an intriguing system for the study of ecological speciation, or the formation of reproductive barriers as a result of divergent, ecologically based selection. To provide phylogenetic context for intraspecific, population-level analysis of patterns of divergence among *Ua* populations, we extend the taxonomic and genetic sampling of previous studies to infer the history of diversification and host-use in the genus. Our scope includes *Ua*'s tribe, the Macrosiphini, which includes the pea aphid (*Acyrthosiphon pisum*) model system. Based on the inferred phylogeny we estimate the timing of divergence events and ancestral shifts in host-use traits among *Uroleucon* species and related lineages.

Macrosiphini

Progress in the systematics and host-use history of the Macrosiphini is of interest because some of its species, particularly *A. pisum*, a genomic model system, provide opportunities to understand how host-use evolution at the population genetic level may contribute to reproductive isolation and diversification at the macroevolutionary level (Godfray, 2010; Peccoud *et al.*, 2009). Macrosiphini is the largest tribe in the aphid sub-family Aphidinae, which comprises a major radiation of aphids that took place during and after the diversification of angiosperms during the Cretaceous and Tertiary, and includes the majority of extant aphid species (Blackman & Eastop, 2008). Aphidinae is typically divided into three tribes—Aphidini and Macrosiphini, which together contain the large majority of species, and the smaller Pterocommatini. Host use is relatively conserved in the Macrosiphini—though they comprise around three-quarters of the described Aphidinae species, macrosiphine taxa collectively utilize fewer plant genera and families than Aphidini (von Dohlen *et al.,* 2006).

Despite its status as a major clade containing the pea aphid, few systematic efforts have been made to revise the tribe. This owes largely to the relatively homogenous morphology observed across its taxa—extant classifications generally make use of combinations of continuous characters rather than discrete synapomorphies (von Dohlen *et al.*, 2006). In this study we test the monophyly of Macrosiphini and investigate its host use history using the most extensive molecular sampling of macrosiphine species to date.

Our interest in the evolutionary history of macrosiphines is primarily to develop *U. ambrosiae* as a system for the genomic study of intraspecific host-use evolution. To achieve this it is important to clarify the evolutionary relationship between *A. pisum* and *Ua*. Pea aphids form genetically divergent host-races—possibly representing incipient species—on different host plants, mostly in the Fabaceae (Hawthorne & Via, 2001; Peccoud *et al.*, 2009; Simon *et al.*, 2003). *Ua* populations, in contrast, use relatively distantly related hosts in the Asteraceae. In general, *Acyrthosiphon* species utilize Fabaceae, Rosaceae and Eurphorbiaceae hosts, while the 160 *Uroleucon* species nearly all utilize Asteraceae (Blackman & Eastop, 2008). Among other causes, then, sequence-level divergence between the pea aphid and *Ua* should reflect the accumulation of neutral differences since they diverged from a common ancestor, as well as diversifying selection driven in part by divergent host use. It is thus of interest to identify the likely host plant used by the common ancestor of *Acyrthosiphon* and *Uroleucon*, as well as to describe any host shifts that may have occurred as the two lineages diverged.

Uroleucon

We test support for the monophyly of *Uroleucon* using a dataset that significantly extends sampling of genetic loci and the number of *Uroleucon* taxa, building on the previous effort by Moran *et al.* (1999). While no evidence presently contradicts the monophyly of the genus, phylogenetic inference in aphids is notoriously difficult (von Dohlen, 2000; Nováková *et al.*, 2013). This condition is prevalent in *Uroleucon*, a group of morphologically and ecologically convergent dark-colored species feeding on various subgroups of host species in the sunflower family (Asteraceae, also known as Compositae) (Moran *et al.*, 1999). As a result, recent attempts to resolve the phylogenetic history of aphids in general and *Uroleucon* in particular have focused on molecular rather than morphological data, in the process overturning many of the higher-level, morphology-based classifications proposed by earlier authors (von Dohlen *et al.,* 2006).

Uroleucon has been subdivided into three subgenera mainly on the basis of three continuous morphological characters, only one of which distinguishes the subgenera *Uroleucon* Mordvilko and *Uromelan* Mordvilko (Moran *et al.*, 1999; Olive, 1965). Both subgenera are distributed throughout the northern hemisphere, with some representation in South America, whereas the subgenus *Lambersius* is restricted to the New World. However a number of species exhibit intermediate traits that are difficult to sort under the current system, and Moran *et al.* (1999)

found that none of the three subgenera is monophyletic, with the possible exception of *Lambersius*. That result was based on sampling only 14 *Uroleucon* species and utilized four loci—three linked mitochondrial and one nuclear locus. Aside from *Ua,* only two other representatives of Nearctic *Uroleucon* were included.

Biogeography and host use

In reconstructing the history of host associations during *Uroleucon* diversification, we focus on three host-use traits. *Host alternation* is an aphid-specific trait involving the obligate use of unrelated plant taxa during sequential seasons, often a woody spring and summer host followed by an herbaceous host in the autumn. *Host choice* refers to the specific host plant taxa utilized by an herbivore lineage. *Host breadth* refers to the range of acceptable hosts on a continuum from extreme specialism, in which a taxon utilizes one or several host plant species in a single genus, to extreme generalism in cases where an insect taxon exploits many phylogenetically and chemically dissimilar host species (Futuyma & Moreno, 1988). Clarifying the ancestral character states for these host-use traits will provide a framework to help evaluate the role potentially played by ecologically based divergent selection (in this case, host-mediated) in the diversification of *Uroleucon* species.

A major aim of this study is to characterize patterns of all three host-use traits in Macrosiphini, and *Uroleucon* in particular. Is there a dominant pattern in the distribution of host-alternation or host-use traits, e.g. phylogenetic niche conservatism? More specifically, we attempt to infer ancestral host use for the most recent common ancestor of *Acyrthosiphon* and *Uroleucon*, and the likely ancestral host taxon of *Ua* and its close relatives. To address these questions, we use phylogenetic mapping and ancestral character state reconstruction to estimate the ancestral host use traits of the common ancestor of *A. pisum* and *Ua,* along with several other ancestral nodes of interest as described below.

The evolution of host choice is likely to be closely associated with the evolution of host breadth. How common are transitions in host breadth, *i.e.* from specialist to generalist or the reverse? Recent theories of herbivorous insect diversification view generalism as a catalyst for diversification by integrating new hosts into the diet, creating the potential for descendent lineages to become reproductively isolated on one or another host lineage (Janz & Nylin, 2008). This theory is consistent with findings that transitions to and from specialism are equally

common, and predict that there are limits to phylogenetic niche conservatism: generalist host-range traits may be expected to be common near the base of clades that have adaptively radiated. Southwestern *Ua* populations are generalists while eastern populations are specialists—a pertinent question is whether *Ua's* closest relatives tend to exhibit specialist or generalist host ranges.

Finally, we estimate divergence times among key macrosiphine lineages to advance two related goals. First, we test the hypothesis suggested by Moran *et al.* (1999) that *Ua* is part of a recent, rapid, adaptive radiation—is there evidence that *Ua* shares relatively shallow divergence times with its closest congeners, relative to other members of the genus? Second, we examine the timing of colonization and diversification events within *Uroleucon* and Macrosiphini in relation to the diversification of host plant lineages.

Approach

To address our questions, we aggregated existing sequence data to reconstruct the phylogeny of genera within the Macrosiphini, with a focus on *Uroleucon.* Relative to the first phylogenetic treatment of *Uroleucon* (Moran *et al*., 1999), we use more dense taxon sampling (35 species), including four times the number of Nearctic species. We also sample a larger and more diverse set of genetic loci, including three mitochondrial, one nuclear, and two loci from the bacterium *Buchnera aphidicola,* an obligate endosymbiont of aphids with utility as a phylogenetic marker due to vertical transmission and a history of co-speciation with aphid host lineages (Clark *et al.*, 2000). Host-use traits mapped to the resulting trees help clarify the host-use transitions involved in an adaptive radiation. We use our phylogenetic analysis to answer the following questions.

First, we test the monophyly of the tribe Macrosiphini, the genus *Uroleucon*, and its three subgenera. Second, we use phylogenetic mapping of host associations to clarify our understanding of ancestral host use traits (*i.e.* host-use, host-breadth, host alternation) for the common ancestor of *Acyrthosiphon* and *Uroleucon*, the common ancestor of *Uroleucon* species, and the close Nearctic relatives of *U. ambrosiae*. Third, we address the biogeographic origin and estimate the timing of diversification of the North American clade that includes *U. ambrosiae*, with the main goal of evaluating the hypothesis that *Ua* originated as part of a recent adaptive radiation.

METHODS

Sequence and character data

Two sequence data sets were assembled for this study. All sequences were obtained from NCBI Genbank (Benson *et al.*, 2010). The "Macrosiphini-dataset" includes the largest set of macrosiphine species for which sufficient sequence data was available. To identify taxa with available homologous sequence data, all available Macrosiphini sequences were downloaded from the Genbank nucleotide sequence database, regardless of sequence identity. Sequences were then clustered using a 0.6 identity threshold using USEARCH (Edgar, 2010). After removing duplicate species entries and manually confirming shared annotations for the sequences in each cluster, loci were selected and grouped such that the resulting dataset maximized the number of Macrosiphini species with sequence data available for at least three out of six target loci. These loci include: cytochrome oxidase I (COI), cytochrome oxidase II (COII), and partial 12S rRNA + tRNA-valine (tRNA-val), all mitochondrial loci; the elongation factor-1 alpha (ef1a) nuclear locus; and two loci from the bacterial endosymbiont *B. aphidicola* (groEL, trpB). Outgroup taxa were selected on the basis of availability of sequence data for all six loci. Our sampling strategy is intended to allow robust phylogenetic inference for the maximum number of species possible, even if only partial data was available for some. Although missing data is not ideal, partial datasets (in some cases with near half missing data) have proven widely useful for phylogenetic inference (Wiens & Tiu, 2012; Wiens & Morrill, 2011). In particular, standard support measures including bootstrap values and Bayesian posterior probabilities retain robust support of nodes even when data sampling is uneven across taxa.

The second set of sequences, the "*Uroleucon*-dataset," is intended to provide the most extensively sampled phylogenetic analysis of *Uroleucon* species to date. It was assembled similarly to the Macrosiphini-dataset as described above but includes the largest number of *Uroleucon* species with at least two sequences available of the following four: mitochondrial (COI), nuclear (ef1a), and two *Buchnera* regions (groEL, trpB).

Sequences were aligned using MUSCLE with default settings as implemented in Geneious (Drummond *et al.*, 2011). Ends and intronic regions were trimmed in Geneious to avoid spurious alignments.

Model selection

Models for each gene were selected using PartitionFinder (Lanfear *et al.*, 2012) with default settings, allowing for unlinked branch lengths among loci, i.e. differing gene histories. Two loci, ef1a (nuclear) and groEL (*Buchnera*), mainly comprised open reading frames, and each codon position was correspondingly partitioned to allow for independent evolutionary models accounting for different mutation rates. The best fit for all loci was the General Time Reversible plus Invariant sites plus Gamma distributed model (GTR+I+G), which was used for maximum likelihood (ML) and Bayesian inference methods.

Phylogenetic analysis

The trimmed Macrosiphini-dataset alignments were used for phylogenetic analysis using maximum parsimony (MP), ML and Bayesian inference. For MP, all loci were concatenated into a single matrix and analyzed using MEGA 5.2 (Tamura *et al.*, 2011) with default settings and 1000 bootstrap replicates. For ML, the data were partitioned by gene and run using raxML (Stamatakis, 2006) with default settings and 1000 bootstrap replicates.

Multi-locus Bayesian analysis for both data sets was conducted with BEAST (Drummond & Rambaut, 2007) using a lognormal relaxed uncorrelated clock model to account for lineage-specific rate heterogeneity. Tree priors for each parameter were specified using the Yule model, a simple model of speciation appropriate for sequences from different species (Drummond & Rambaut, 2007).

Divergence estimation

Divergence times were estimated using a multilocus Bayesian MCMC approach implemented using the BEAST software package v1.8.0 (Drummond *et al.*, 2012). BEAST estimates divergence times based on molecular clock analyses without conditioning on a single tree topology. The analysis pipeline followed recommendations of the authors, including use of BEAUti v1.8.0 to construct an input file, Tracer v1.6 to examine convergence of MCMC chains, and LogCombiner v1.8.0 to combine runs (Drummond & Rambaut, 2007). Clock and substitution rate models were as described above for phylogenetic analysis. Final results are based on two MCMC runs with 100 million generations, sampled every 1000 generations. The first million generations were ignored as burn-in. Mean parameter estimates and support values

(95% highest posterior densities) were summarized with TreeAnnotator v1.8.0. Trees were visualized using FigTree v1.4.0 (Drummond & Rambaut, 2007).

Calibration points for the molecular clock estimates were derived from several recent studies integrating molecular data with the sparse aphid fossil record (von Dohlen, 2000; von Dohlen *et al*., 2006; Kim *et al.*, 2011; Moran *et al.*, 1999). The common ancestor of Macrosiphini genera was constrained to a minimum age of 43 MYA and a maximum age of 53 MYA, and the common ancestor of Macrosiphini and Aphidinae was calibrated to 60 +/- 5 MYA. These estimates follow those used by Kim *et al.* (2011) based on fossil calibrations of the Aphididae and Aphidoidea crown clades 150 and 90 MYA, respectively.

Ancestral trait reconstruction

Ancestral states were reconstructed for three characters: host alternation and host choice for the Macrosiphini-dataset, and biogeographic origin for the *Uroleucon*-dataset. Reconstructions were inferred according to Bayesian criterion using the BayesMultiState method of the BayesTraits version 2.0 package (Pagel & Meade, 2007; Pagel *et al.*, 2004). To minimize the effect of uncertain priors, the reverse jump hyperprior approach was used, as recommended, to generate a distribution of priors. For each run, a range of hyperprior values and burn-in settings were used to find MCMC acceptance rates between 20-40% (Pagel & Meade, 2007). Ultimately, the hyperprior range was set from 0-30, and rate deviation was set to autotune. BayesTraits accounts for uncertainty in tree topology and branch length; the ten trees with the highest cumulative posterior probability as calculated by BEAST were used for each dataset. Each analysis was run for one million generations and sampled every 1,000 generations, following a burn-in of 50,000 generations. The mean for each prior was calculated for generations after which the MCMC chain became stationary.

To reconstruct ancestral host alternation, taxa were assigned as either monoecious holocyclic (simple use of an herbaceous plant) or heteroecious (host-alternating), based on the presence of the trait within their genus (Blackman & Eastop, 2008). Taxa with facultative host-alternation were treated as heteroecious as long as host-alternation has been recorded for some members of the genus. To reconstruct ancestral biogeographic origin, each taxon was assigned to one of four states, based on data given by Blackman & Eastop (2008): a) North America, b) Europe (extending to western Asia), c) Asia (extending to eastern Europe), or d) holarctic. For

the Macrosiphini-dataset, host use records were used to assign each species to one of the following states: a) single rosid family (29.4% of the 68 species), b) single asterid family (47%), c) generalist, for species recorded on at least three plant families (17.6%), d) other (for taxa that utilize fewer than three plant families that are neither asterids or rosids; 6 taxa or 17.6%). These data are based on the number of host plant genera reported for each species in Blackman & Eastop (2008).

Relative probabilities of ancestral character states are based on the posterior densities of each reconstructed state and the posterior probability of the node, accounting at once for uncertainty about the ancestral state and uncertainty about the node itself (Pagel *et al*., 2004).

RESULTS

Phylogenetic relationships

The MP, ML and Bayesian analyses of the Macrosiphini-dataset broadly agree on the monophyly of major groups of genera—no clade with high support in one tree is contradicted by a well-supported alternate grouping in another tree (Figures 1-2). Although taxa for this study were not selected expressly to test higher-level relationships, the results are consistent with previous studies. Macrosiphini is supported as a monophyletic group in all analyses, with the exception of two taxa that fall outside the group (Figure 2). The two, *Cavariella* and *Capitophorus*, are allied with Pterocommatini, forming the sister group to Macrosiphini+Aphidini. This arrangement contradicts present nomenclature (Blackman & Eastop, 2008), but is consistent with previous molecular analyses that exclude one or both genera from Macrosiphini (Kim *et al.*, 2011; Nováková *et al.*, 2013; Papasotiropoulos & Tsiamis, 2013). *Cavariella* and *Pterocomma* both feed on Salicaceae, suggesting that, in this case, conserved host-use traits may be more reflective of shared ancestry than is morphology, which is the basis for their present classification as unrelated taxa (Peccoud *et al.* 2010).

Existing molecular phylogenetic hypotheses disagree on the relationships of the three tribes of Aphidinae, i.e. Aphidini, Pterocommatini, and Macrosiphini. The traditional classification places Pterocommatini basal to the other two tribes, and is supported by several recent studies (Kim *et al*., 2011; Nováková *et al.*, 2013; Ortiz-Rivas & Martínez-Torres, 2010; Papasotiropoulos & Tsiamis, 2013). von Dohlen *et al.* (1999) describe evidence supporting a basal placement of Aphidini. Our analysis instead provides support for Pterocommatini as the sister group to Macrosiphini+Aphidini, two of the largest and most diverse aphid groups. Support for this arrangement comes mainly from the Bayesian analysis; in the MP and ML analyses Macrosiphini forms an unresolved polytomy with Aphidini and Pterocommatini. This is generally congruent with previous molecular phylogenetic analyses of these groups, which all found support for clades up to the sub-tribal level, but provide contradicting hypotheses for higher-level relationships (von Dohlen, 2000; von Dohlen *et al*., 2006; Kim *et al.*, 2011; Moran *et al*., 1999; Nováková *et al.*, 2013; Ortiz-Rivas & Martínez-Torres, 2010; Papasotiropoulos & Tsiamis, 2013). We also replicate previous work in finding that *Myzus* and *Dysaphis* are polyphyletic—these groups are known to be in need of revision. Sparse sampling of sub-tribal

diversity is likely a contributing factor to these issues. The Aphidini includes ~750 species, and of these, no molecular analysis, including the present one, has considered more than 8 species in 4 genera (von Dohlen *et al.*, 2006; Kim *et al.*, 2011; Nováková *et al*., 2013; Ortiz-Rivas & Martínez-Torres, 2010; Papasotiropoulos & Tsiamis, 2013). More extensive sampling of Aphidinae sub-tribal lineages is likely to be critical to improving phylogenetic resolution of relationships among this major group of aphids.

Uroleucon monophyly

To reconstruct the host-use history of *Uroleucon* species we build on a previous molecular analysis of fourteen species in the genus by Moran *et al.* (1999). In addition to the Macrosiphini-dataset, which includes 20 *Uroleucon* species, we assembled a second dataset, the *Uroleucon*-dataset, with 35 *Uroleucon* species selected based on availability of sequence data at four loci. Bayesian analysis of the Uroleucon-dataset was used to co-estimate phylogeny and associated divergence times calibrated with dates based on fossil data. Each dataset yielded consistent topologies for all overlapping taxa. While relationships among many of the additional taxa in the *Uroleucon*-dataset are unresolved, the topology provides novel insight into divergence times and host-use evolution of major clades in the genus. The majority of the newly analyzed taxa are associated either with a closely related group that includes *U. ambrosiae,* or a paraphyletic grade with deep branch lengths at the base of the tree. This discussion refers only to inferred relationships with posterior probability support of at least 50%.

The monophyly of *Uroleucon* is well supported in all analyses of both datasets (Macrosiphini-dataset, Figure 2; *Uroleucon*-dataset, Figure 3). Not one of the three *Uroleucon* subgenera—*Uroleucon*, *Uromelan*, and, more tentatively, *Lambersius*—is monophyletic, consistent with the determination of Moran *et al.* (1999) that *Uroleucon* and *Uromelan* are paraphyletic. Deep branch lengths separating species in the strictly Nearctic sub-genus *Lambersius* suggest that it contains the oldest lineages in the genus, pointing to a North American origin. Ancestral character state reconstruction of biogeographic origin also supports a North American origin for the genus. This conclusion, however, remains tentative as no South American taxa were included in the present analysis.

A core group of Nearctic species in the subgenus *Uroleucon*, including *U. ambrosiae*, form a well-supported monophyletic group of closely related species (*Ua* clade, Figures 4-5).

Monophyly for all Nearctic *Uroleucon* is disrupted by the appearance of a single *Uromelan* species, *U. eupatorifoliae*. Because this species is the only Nearctic *Uromelan* under study, it is difficult to assess whether it is representative of a larger derived Nearctic *Uromelan* clade. Accounting for this uncertainty, Nearctic *Uroleucon* includes one, or two closely related, clades. Under either scenario, Nearctic *Uroleucon* species appear to be sister to and likely descended from *Uromelan* species of Palearctic origin following a single colonization event of North America (*U. aeneum* clade, Figure 5).

The remaining members of subgenus *Uroleucon* included in this study, all Palearctic, are placed basal to Nearctic *Uroleucon* + Old World *Uromelan*. The Nearctic clade of *Uromelan* (*U. rurale* + *U. helianthicola*) appears to be relatively old and possibly share a common ancestor with *Lambersius* species.

Divergence timing

Estimated divergence times for ten *Uroleucon* nodes with high posterior support are given in Figures 4-5. Highest posterior density (HPD) intervals, a Bayesian analog of confidence intervals (Drummond & Rambaut, 2007), are provided in brackets. We estimate that *Uroleucon* originated 17.1 – 20.9 million years ago. This is considerably earlier than the estimate of 5 million years calculated by Moran *et al.* (1999). That estimate was based on a constant mitochondrial molecular clock model applied uniformly across fourteen species (Moran *et al.*, 1999). In favor of the present divergence time estimate, Bayesian inference methods allow branch-specific probabilistic modeling of molecular evolution rates, which, especially when applied to unlinked loci, are expected to be preferable to approaches based on a strict clock (Drummond *et al.*, 2012).

Uroleucon lineages almost certainly colonized North America at least twice, most recently around 3.7 MYA, during the late Tertiary (Figure 5). This colonization event appears to have resulted in two major North American lineages. The first, including *U. rudbeckiae*, may have diversified shortly after colonization (around 2.7 MYA). The second comprises ten species in the analysis, including *U. ambrosiae,* that all diversified much later, in the Pleistocene, around one million years ago (HPD interval 0.5 – 1.9 MYA).

Ancestral host use and biogeography

Phylogenetic character mapping and ancestral character state reconstruction were carried out to help infer key processes in *Uroleucon* host use history and diversification. The phylogenetic distribution of host use traits (Figure 6) illustrates that aphidine lineages have diversified to feed on a variety of angiosperm families. Consideration of the higher-level relationships of these host plant taxa reveals a degree of host-use conservation. Three angiosperm clades—asterids, rosids, and grasses (Poaceae)—comprise the majority of host plant taxa used by the lineages under consideration. Despite apparent conservation of rosid host use among early lineages, host shifts have occurred in all tribes, including at least three shifts to asterids, and at least two shifts to generalist host-breadth. Host use within the Macrosiphini mirrors host use in the Aphidinae as a whole, with many of the same families appearing repeatedly, especially Asteraceae, Rosaceae and Fabaceae, with repeated transitions among these and other asterid and rosid families.

To strengthen conclusions regarding ancestral host use traits, character states were reconstructed accounting for uncertainty in the topology as well as the appropriate models for evolution for the character state transitions in question. Reconstructed characters included host use (of higher-level host clades, *i.e.* asterid, rosid, etc.), biogeographic origin, and host alternation.

Results for ancestral host use suggest that Aphidinae and Macrosiphini most likely originated feeding on rosids (relatively probability 76% for each), as did the common ancestor of *Uroleucon* and *Acyrthosiphon* (U+A clade in Figure 6). These findings support the conclusion that early lineages diversified on rosids, giving rise to some descendents that shifted to asterids and Poaceae, and others that evolved into some of the most extensively generalist aphid genera (*e.g. Myzus* and *Dysaphis*). In contrast to earlier diverging lineages, the common ancestor of *Uroleucon* and *Macrosiphoniella* utilized Asteraceae with 99% probability.

Both *Uroleucon* and its sister genus, *Macrosiphoniella*, use hosts in the Asteraceae to the near exclusion of other families. Within *Uroleucon*, both basal and more recently diverged lineages (*e.g.* the Ua clade in Figure 4) show conserved use of one Asteraceae sub-family, Asteroideae. However the latter likely represents an independent colonization, as suggested by the use of hosts in other Asteraceae sub-families (Cichoroideae and Carduoideae) and the family Campanulaceae, by lineages intermediately positioned in the inferred topology (Figure 4). Along with transitions in host-use, *Uroleucon* taxa exhibit varying levels of host-breadth. To create a

visual representation of the distribution of host-breadth traits, bar graphs representing the number of distinct host genera utilized by each species are given in Figure 4. Like most aphids, *Uroleucon* species are predominantly specialists. Yet repeated transitions to more generalized host-breadth are evident in several lineages, including those containing *U. ambrosiae, U. sonchi,* and *U. erigeronensis*.

Reconstruction of ancestral biogeographic distributions was focused on *Uroleucon* nodes. At 55% probability, a North American origin for the genus is equivocal. Poor resolution of relationships among basal *Uroleucon* lineages is one likely reason for uncertainty in reconstructing this node. In contrast, there is high likelihood that *Uroleucon* species have radiated in North America following a re-colonization of the continent between around 3.7 – 6.5 MYA (Figure 5).

Ancestral reconstruction of host-alternation traits supports the now prevailing view that host-alternation has been readily gained and lost in the course of aphid evolution (Figure 7) (von Dohlen, 2000; von Dohlen *et al.*, 2006). Evidence is lacking for ancestral host-alternation at the root of Aphidinae, the root of Macrosiphini, or in the common ancestor of *Uroleucon* and *Acyrthosiphon* (Figure 7). Nevertheless, host alternation is evident in numerous extant taxa across Aphidinae lineages, implying multiple losses and acquisitions.

The *Ua* and *A. pisum* clades show divergent patterns in host alternation. The most recent common ancestor of *Uroleucon* very likely fed obligately on herbaceous asterid species (Figure 6). This conclusion squares with the observation that host alternation is entirely absent in the hundreds of species contained in *Uroleucon* and *Macrosiphoniella* (Blackman & Eastop, 2008). Host alternation is also absent in *Metopeurum*, a genus previously associated with *Macrosiphoniella* (but not included here due to lack of sufficient sequence data) which also feeds on Asteraceae species (von Dohlen *et al.*, 2006). In contrast, host-alternation is evident in several genera within the *A. pisum* clade, though not in *Acyrthosiphon* itself.

DISCUSSION

Ancestral host use patterns in the Aphidinae

The results from phylogenetic mapping and ancestral character reconstruction of host use traits provide the outlines of key events in the host-use history and evolution of *Uroleucon* aphids, from the divergence of the Macrosiphini from other aphid tribes around 60-65 MYA (Kim *et al.*, 2011), to an adaptive radiation of the *U. ambrosiae* group on various Compositae lineages in the last 2 million years.

We find that the common ancestor of extant Macrosiphini, which lived around 48 MYA, likely fed on rosid host plants. This ancestor appeared 10-15 MYA later than the common ancestor of Macrosiphini and Aphidini (Kim *et al.*, 2011), which probably also utilized related hosts. Rosids, comprising nearly a fourth of extant angiosperm taxa, were already diverse at this time, following a rapid radiation 83-108 MYA (Wang *et al.,* 2009). This timing supports a scenario in which aphidine diversification lagged behind that of its primary hosts.

A subsequent radiation is thought to have taken place in the mid-Tertiary (Miocene or later), around 23-33 MYA, when climatic cooling favored the spread of herbaceous angiosperms like composites and grasses (von Dohlen, 2000; Heie, 1987). Several adaptive radiations likely took place around this time in lineages utilizing herbaceous angiosperms as primary or secondary hosts (von Dohlen, 2000). Given the origin of *Uroleucon* by 20 MYA, it is likely that the origins of the major sub-tribal macrosiphine lineages date to this period. This radiation took place around the time the Asteroideae originated 27-30 MYA, and preceded the radiation of Heliantheae, Astereae and other common hosts of extant *Uroleucon* species.

One hypothesized consequence of Tertiary cooling is the expansion of temperate zones, which may have favored the ability to feed on woody hosts, as most host-alternating species do. Host-alternation is thought to be an adaptation to seasonality, and a reason why aphid diversity is concentrated in temperate latitudes, an unusual pattern relative to other insect groups that are most diverse in the tropics (von Dohlen, 2000). While many distantly related Aphidinae lineages exhibit host-alternation, we find scant support that it was an ancestral trait for Macrosiphines or the common ancestor of *Uroleucon* and *Acyrthosiphon* (Figure 7). These findings comport with recent phylogenetic treatments that support the view that host-alternation arose multiple times in the course of aphid diversification during the early Tertiary or late Cretaceous (50-70 million

years ago) (von Dohlen, 2000; von Dohlen *et al.*, 2006; Peccoud *et al*., 2010). The evolutionary lability of host alternation appears to have continued during subsequent radiations that gave rise to present-day macrosiphine diversity.

While our study is commensurate in sampling scope to existing analyses, it is worth noting that our conclusions are based on relatively sparse sampling of basal Aphidine lineages (e.g. Aphidini). Poor resolution among basal branches is a likely consequence, which adds to the uncertainty of ancestral character state reconstruction. A reconstruction method that accounts for uncertainty in the topology by considering multiple trees helps to account for this. Regardless, more extensively sampled phylogenetic studies are needed to deepen our understanding of this phase of aphid evolution.

Uroleucon *biogeography and host use evolution*

The focus of this analysis is to clarify the biogeography and host-use history of *Uroleucon,* with a focus on *U. ambrosiae*. Most prominently, we find that *Ua* is part of a large adaptive radiation of largely sympatrically distributed species that have diversified from a common ancestor in the last 1-2 million years.

We find that *Uroleucon* originated around 20 MYA. This timing places the early diversification of *Uroleucon* in the Miocene, coincident with the appearance of the major Asteraceae tribes in the fossil record (Heie, 1996; Moran *et al*., 1999). This period saw continued cooling, expansion of temperate zones, and the establishment of grasslands dominated by grasses and herbaceous angiosperms, including a rapidly diversifying number of Asteraceae species. The host lineages most commonly exploited by *Uroleucon* species, Astereae and Heliantheae, diversified around 17 MYA (Funk & Oberprieler, 2009; Pelser & Watson, 2009). The earliest diverging *Uroleucon* lineages (*Lambersius* and Nearctic *Uromelan*) fed on what were likely basal Astereae and Heliantheae species.

Poor resolution of divergences that occurred between 18.9 and 8.8 million years ago make it difficult to detail that period. However, given a probable North American origin as well as the cluster of European and Asian species (Palearctic *Uroleucon*) separating North American clades, one or more of these early species colonized Eurasia (Figure 5). Two routes are most likely: to Europe over the North Atlantic land bridge, which persisted until around 15 MYA or earlier, or to Asia via the Bering land bridge, which was open throughout the late Tertiary until

around 5.5 million years ago (Milne, 2006). Angiosperm distributions are known to follow disjunct distributions resulting from the submersion of both bridges, pointing to the possibility that *Uroleucon* migrants may have been tracking their host plants.

Transition to Europe was followed closely by host-shifts to species in the Compositae sub-families Cichorioideae and Carduoideae. Both families are relatively distant from the ancestral host tribes Astereae and Heliantheae (Asteroideae), and exhibit low diversity in North America. This is a marked pattern, as all North American *Uroleucon* species in the analysis feed exclusively on Asteroid hosts—all non-Asteroid host-use occurs in Eurasia, where these two subfamilies are much more diverse (Figures 4,5). The only exceptions are records for non-Asteroid hosts among generalist populations of *U. ambrosiae* in southwestern North America—a capacity that may potentially be ancestrally retained, in light of these results. At least two *Uroleucon* species, one in Europe and another in Asia, shifted away from Compositae, feeding instead on Campanulaceae species.

One or more Asian *Uromelan* species re-colonized North America between 3.7 − 6.5 MYA. The most likely route is the Bering land bridge, putting the date no later than 5.5 MYA, when it was submerged until the Pleistocene (Milne, 2006). This geographic shift was accompanied by a host shift back to Asteroideae, setting the stage for an adaptive radiation in the *Ua* clade on Compositae species beginning 0.5-2.0 MYA. This timing is coincident with the Pleistocene Last Glacial Maximum, when many North American taxa were driven into southerly refugia. As the glaciers receded, Compositae species and other plants expanded geographically and phylogenetically, providing an opportunity for aphids to expand their range and encounter novel associations of hosts (Funk & Oberprieler, 2009).

Several lines of evidence together provide strong support for an adaptive radiation of the *Ua* clade. In all phylogenetic analyses of both datasets in this study, members of the clade all have very short branch lengths. This is despite variation at the genetic loci sampled, which gave rise to the recent estimated divergence time of the common ancestor of the clade. In addition, nearly all species in the clade have an overlapping range in eastern North America. Finally, species in the group exploit relatively closely related host plants, all of which are in derived composite lineages.

Most *Uroleucon*, like most aphids and most insect herbivores, exhibit relatively specialized host-breadth. While most species in the *Ua* clade conform to this pattern, several exploit a larger

number of host plant genera (Figure 4). Further sampling and robust reconstruction of host-breadth will be required to begin to understand the dynamics underlying transitions among specialization and generalism. Yet the host-use of the group as presently understood suggests a speculative scenario. *Uroleucon* species have repeatedly and apparently independently evolved to feed on two tribes, Heliantheae and especially Astereae. In the case of Heliantheae, each species utilizes a distinct genus in the tribe. Each species that targets Astereae uses a different species, or set of species, in a single genus, *Solidago*. It may be the case that host associations in the *Ua* clade reflect a sorting process in which related lineages specialize on related but distinct host species. In this scenario, the presence of generalists among the specialized species raises the possibility that genetic variation for ancestral host use traits are retained in some lineages, potentially serving as intermittent sources of populations that evolve specialized host use.

The role of ecological divergence—such as that caused by divergent host use—in initiating and deepening reproductive barriers is a major topic in evolutionary ecology. The *Ua* clade provides an opportune system to understand the relative importance of two different evolutionary processes, which may not be mutually exclusive. First, divergent selection imposed by assortative host-specialization among populations may contribute to reproductive barriers in sympatry due, for example, to hybrid depression or host-associated assortative mating. Second, genomic differentiation among populations may have accumulated by isolation by distance in Pleistocene glacial refugia, reducing the genomic compatibilities of populations once they were back in contact. In this scenario, assortative mate use may result from competitive exclusion.

A related question concerns the genomic basis of host use. Basal *Uroleucon* (i.e. *Lambersius*, Nearctic *Uromelan*) utilize Astereae and Heliantheae species. Their European descendants switched to different hosts, until ~15 million years later, when, upon re-colonization of North America, descendent lineages again shifted to the hosts of their distant ancestors. Do distantly related *Uroleucon* clades that exhibit convergent host use traits share conserved genotypes at the functional or genomic levels?

These findings contextualize host-breadth divergence among *U. ambrosiae* populations and suggest avenues for further research. It is unclear, for example, whether generalized host use is the ancestral or derived state. Finding generalism to be the ancestral trait would support the hypothesis that the evolution of generalism promotes diversification by allowing host-specific diversification via specialization, such as exists throughout the rest of the *Ua* clade. Comparative

genomics approaches to identify candidate loci that compare variation in host-use associated genes—like proteases, metabolic enzymes and effectors responding to phloem-mediated immune and defensive pathways—with neutral loci would be well-suited to address these questions.

CHAPTER 2 REFERENCES

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2010). GenBank. Nucleic Acids Research, *38* (Database issue), D46–D51.

Bernays, E. A., & Funk, D. J. (2000). Electrical penetration graph analysis reveals population differentiation of host-plant probing behaviors within the aphid species *Uroleucon ambrosiae*. Entomologia Experimentalis et Applicata, *97*(2), 183–191.

Bernays, E., Funk, D., & Moran, N. (2000). Intraspecific differences in olfactory sensilla in relation to diet breadth in *Uroleucon ambrosiae*. Journal of Morphology. 245, 99-109.

Blackman, R. L., & Eastop, V. F. (2008). *Aphids on the World's Herbaceous Plants and Shrubs*. John Wiley & Sons.

Clark, M., Moran, N., Baumann, P., & Wernegreen, J. (2000). Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. Evolution, *54*(2), 517–525.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, *7*(1), 214.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Wilson, A. (2011). Geneious v5. 4.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, *29*(8), 1969–1973

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, *26*(19), 2460–2461.

Funk, D., & Bernays, E. (2001). Geographic variation in host specificity reveals host range evolution in *Uroleucon ambrosiae* aphids, Ecology, 82(3), 726-739.

Funk, V. A., & Oberprieler, C. (2009). Compositae metatrees: the next generation. In: Funk,Vicki A., (ed.) Systematics, evolution and biogeography of Compositae. International Association for Plant Taxonomy. 747-777.

Funk, V. A. (ed.) (2009). Systematics, evolution, and biogeography of Compositae. International Association for Plant Taxonomy.

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. Annual Review of Ecology and Systematics, 19(1), 207–233.

Futuyma, D. J., & Mitter, C. (1996). Insect-plant Interactions: The evolution of component communities. Philosophical Transactions of the Royal Society B: Biological Sciences, *351*(1345), 1361–1366.

Futuyma, D. J., & Agrawal, A. (2009). Macroevolution and the biological diversity of plants and herbivores. Proceedings of the National Academy of Sciences, 106(43), 18054-18061.

Godfray, H. C. J. (2010). The pea aphid genome. Insect Molecular Biology, 19(S2), 1–4.

Hawthorne, D., & Via, S. (2001). Genetic linkage of ecological specialization and reproductive isolation in pea aphids. Nature, 412(6850), 904–907.

Heie, O. (1987). Paleontology and phylogeny. In A. K. Minks & P. Harrewijn (Eds.), Aphids: Their Biology, Natural Enemies and Control. Amsterdam: Elsevier.

Heie, O. (1996). The evolutionary history of aphids and a hypothesis on the coevolution of aphids and plants. Bollettino Di Zoologia Agraria E Di Bachicoltura, 28, 149–155.

Janz, N., & Nylin, S. (2008). The oscillation hypothesis of host-plant range and speciation. In: K.J. Tilmon (Ed.), Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects. University of California Press, Berkeley, 203–215.

Kelley, S. T., & Farrell, B. D. (1998). Is specialization a dead end? The phylogeny of host use in *Dendroctonus* bark beetles (Scolytidae). Evolution, 52(6), 1731–1743.

Kim, H., & Lee, S. (2008). A molecular phylogeny of the tribe Aphidini (Insecta: Hemiptera: Aphididae) based on the mitochondrial tRNA/COII, 12S/16S and the nuclear EF1α genes. Systematic Entomology, *33*(4), 711–721.

Kim, H., Lee, S., & Jang, Y. (2011). Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): Diversification, host association, and biogeographic origins. PLoS ONE, *6*(9), e24749.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29(6), 1695–1701.

Milne, R.I. (2006). Northern Hemisphere plant disjunctions: a window on Tertiary land bridges and climate change? Annals of Botany, *98*(3), 465–472.

Mitter, C. C., Farrell, B. B., & Futuyma, D. J. D. (1991). Phylogenetic studies of insect-plant interactions: Insights into the genesis of diversity. Trends in Ecology & Evolution, 6(9), 290–293.

Mitter, C., Farrell, B., & Wiegmann, B. (1988). The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? American Naturalist, 132(1), 107.

Moran, N. (1984). The genus *Uroleucon* (Homoptera: Aphididae) in Michigan: key, host records, biological notes, and descriptions of three new species. Journal of the Kansas Entomological Society, 57(4), 596–616.

Moran, N, Kaplan, M., Gelsey, M., & Murphy, T. (1999). Phylogenetics and evolution of the aphid genus Uroleucon based on mitochondrial and nuclear DNA sequences. Systematic Entomology. *24*, 85-93.

Nosil, P. (2002). Transition rates between specialization and generalization in phytophagous insects. Evolution, 56(8), 1701–1706.

Nováková, E., Hypša, V., Klein, J., Foottit, R. G., von Dohlen, C. D., & Moran, N. A. (2013). Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola. Molecular Phylogenetics and Evolution, 68(1), 42–54.

Olive, A. T. (1965). A new subgenus and two new species of *Dactynotus* (Homoptera: Aphididae). Annals of the Entomological Society of America, 58(3), 284-289.

Ortiz-Rivas, B., & Martínez-Torres, D. (2010). Combination of molecular data support the existence of three main lineages in the phylogeny of aphids (Hemiptera: Aphididae) and the

basal position of the subfamily Lachninae. Molecular Phylogenetics and Evolution, 55(1), 305–317.

Pagel, M., & Meade, A. (2007). BayesTraits. Software Distributed by the Author. http://www.evolution.reading.ac.uk/BayesTraits.html

Pagel, M., Meade, A., & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*. 53(5), 673-684.

Papasotiropoulos, V., & Tsiamis, G. (2013). A molecular phylogenetic study of aphids (Hemiptera: Aphididae) based on mitochondrial DNA sequence analysis. Journal of Biological Research – Thessaloniki, 20, 1-13.

Peccoud, J., Ollivier, A., Plantegenest, M., & Simon, J. (2009). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proceedings of the National Academy of Sciences, 106(18), 7495-7500.

Peccoud, J., Simon, J. von Dohlen, C., Coeur d'acier, A., Plantegenest, M., Vanlerberghe-Masutti, F., & Jousselin, E. (2010). Evolutionary history of aphid-plant associations and their role in aphid diversification. Comptes Rendus Biologies, 333(6-7), 474–487.

Pelser, P. B., & Watson, L. E. (2009). Introduction to Asteroideae. In: Funk, V. A., (ed.) Systematics, evolution and biogeography of Compositae. International Association for Plant Taxonomy, 495-502.

Robinson, A. G. (1985). Annotated list of *Uroleucon* (*Uroleucon*, *Uromelan*, *Satula*)(Homoptera: Aphididae) of America north of Mexico, with keys and descriptions of new species. The Canadian Entomologist, 117(08), 1029–1054.

Robinson, A. G. (1986). Annotated list of *Uroleucon* (*Lambersius*) (Homoptera: Aphididae) of America north of Mexico, with a key and descriptions of new species. The Canadian Entomologist, 118(6), 559–576.

Simon, J. C., Carre, S., Boutin, M., Prunier-Leterme, N., Sabater-Munoz, B., Latorre, A., Bournoville, R. (2003). Host-based divergence in populations of the pea aphid: insights from nuclear markers and the prevalence of facultative symbionts. Proceedings of the Royal Society B: Biological Sciences, 270(1525), 1703–1712.

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M. (2008). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. Botanical Journal of the Linnean Society, 133(4), 381–461.

Soltis, D. E., Soltis, P. S., Endress, P. K., Chase, M. W., Soltis, D. E., Soltis, P. S. (2005). Phylogeny and evolution of angiosperms. Sinauer Associates Incorporated.

Soltis, P. S., Soltis, D. E., & Chase, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature, 402(6760), 402–404.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), 2688–2690.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution, 28(10), 2731–2739.

von Dohlen, C. (2000). Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. Biological Journal of the Linnean Society, 71(4), 689–717.

von Dohlen, von, C., Rowe, C., & Heie, O. (2006). A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. Molecular Phylogenetics and Evolution, 38(2), 316–329.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D., Brockington, S. F., Alexandre, R. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. Proceedings of the National Academy of Sciences, 106(10), 3853–3858.

Wiens, J. J. J., & Tiu, J. J. (2012). Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS ONE, *7*(8), e42925.

Wiens, J. J., & Morrill, M. C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Systematic Biology, 60(5), 719–731.

Winkler, I., & Mitter, C. (2008). The phylogenetic dimension of insect-plant interactions: a review of recent evidence. In K.J. Tilmon (Ed.), Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects. University of California Press, Berkeley, 240–263.

Chapter 3.

Comparative genomics of the feeding transcriptome in the aphid *Uroleucon ambrosiae.*

INTRODUCTION

Understanding the genomic basis of adaptive diversification is a major theme in evolutionary biology. Aphids represent promising model systems for the study of functional diversification. They exhibit several traits—including cyclical parthenogenesis, host alternation, wing polyphenism, and varying degrees of host specificity—that point to a key role for host plant use in functional diversification (Brisson & Stern, 2006; Jean & Jean-Christophe, 2010). The pea aphid, *Acyrthosiphon pisum,* is also a model system for understanding the evolution and functional genetics of ecological specialization and the role of host associations in ecological speciation (Hawthorne & Via, 2001; Smadja *et al.*, 2009; Via & Hawthorne, 2002).

With the release of the pea aphid genome in 2010 (IAGC, 2010), genome-scale studies are increasingly feasible in non-model aphid systems. In this study we sequence and annotate the feeding transcriptome of the brown ambrosia aphid, *Uroleucon ambrosiae (Ua)*. We then make use of existing genomic resources in insects, principally the pea aphid genome, to a) identify *Uroleucon* homologs of known host-associated gene families, and b) characterize gene family expansions that may reflect lineage-specific adaptive evolution.

These data will provide a framework for investigating the genomic signature of divergent host-use traits among *Ua* populations, which range from generalist to specialist in their host-breadth. Among aphids for which genomic resources (e.g. EST data sets) are available, the transcriptome data reported here represent the most phylogenetically proximate genomic resource relative to *Acyrthosiphon*—both are members of the dactynotines, one of three major sub-lineages in the aphid tribe Macrosiphini (von Dohlen *et al.*, 2006).

The ancestral lineages of *A. pisum* and *Ua* probably diverged between 25-45 million years ago (MYA). This rough estimate is based on the fact that their common ancestor must have evolved following the origin of Macrosiphini 48 MYA (Kim *et al.*, 2011) and had to have diverged well before the origin of *Uroleucon* around 20 MYA (Gill, Chapter 1). In the time since their divergence, each lineage has adapted to different host-use ecologies. Most *Uroleucon*

species, including *Ua*, specialize on individual species or genera in the Asteraceae or Campanulaceae, and no *Uroleucon* species exhibit host alternation (Moran *et al*, 1999, von Dohlen *et al.* 2006). This is in contrast to many other macrosiphines, which tend to feed on species in the Rosaceae, in some cases as a secondary host for host-alternating species. *A. pisum* is a non-host alternating species that comprises a series of genetically differentiated host races specializing on various legume (Fabaceae) species (Ferrari *et al.*, 2008; Frantz *et al.*, 2006).

Among other changes, genomic divergence between the *Acyrthosiphon* and *Uroleucon* lineages likely involved evolution among gene families that respond to host plant cues. One goal of this study is to identify *Uroleucon* homologs of these gene families, which we summarize below.

Chemosensory loci

Chemosensory loci—including olfactory, gustatory and ionotropic receptor proteins as well as olfactory binding proteins—are of particular interest, as they mediate host plant selection (Bernays & Chapman, 1994) and are known to evolve rapidly at the sequence and expression levels under selection from novel host ecologies (Kopp *et al.,* 2008; McBride, 2007). In the pea aphid, the handful of SNPs that showed the strongest responses to divergent selection among host races nearly all corresponded to olfactory or gustatory genes (Smadja *et al*., 2012). The specialist *Drosophila sechellia*, relative to its more generalist sister species *D. simulans,* shows rapid evolution of olfaction-related gene families, including olfactory binding proteins and olfactory and gustatory receptor gene families, primarily characterized by extensive gene loss and signs of accelerated evolution resulting from selection (Dworkin & Jones, 2008; McBride, 2007; Vieira & Rozas, 2011).

Effectors and salivary proteins

Mounting evidence suggests that effector molecules may play central roles in mediating insect-plant arms races. Aphid stylets penetrate the plant vascular system, requiring interaction with the internal environment of the plant, including multiple layers of the same immune-related pathways that respond to fungal pathogens, for example. Effectors are broadly defined as any insect proteins (or other functional molecules like microRNAs) that alter host cell structure or function (Hogenhout & Bos, 2011). They include diverse gene families that variously modify

induced and systemic plant responses, including defensive responses, in addition to nutrient allocation, secondary chemical synthesis, morphological structure (*e.g.* manipulation of stomatal openings) and other plant properties (Giordanengo *et al.,* 2010; Hogenhout & Bos, 2011; Hogenhout *et al.*, 2009; Thompson, 2006). Ion channel proteins, for example, are effectors known to prevent occlusion in sieve cells. Secreted salivary proteins are involved in a variety of molecular insect-host interactions beginning with the initiation of probing. The pea aphid salivary secretome has been predicted (Carolan *et al.,* 2011; Carolan *et al.*, 2009), with research focusing on one particular gene, C002, shown to be vital to host feeding (Mutti *et al.,* 2008). C002 is highly diverged among aphids ($D_N/D_S = 0.73$) (Ollivier *et al.*, 2010).

Cytochrome p450s represent another class of well-described effectors involved in detoxification of plant secondary defensive compounds (Berenbaum, 2002), and are known to show greater divergence than other genes when related species with differing host use are compared. The generalist aphid *Myzus persicae* has 40% more p450 genes than the pea aphid (Ramsey *et al.,* 2010), perhaps in association with a greater diversity of secondary compounds encountered in its generalist diet. We also examine circadian clock genes, which show accelerated rates of evolution in *A. pisum* relative to other gene families (Cortés *et al.*, 2010). Although it is not clear that circadian clock genes are directly responsive to host plant cues, their possible role in pea aphid diversification flags them as genes of interest.

Lineage-specific gene expansion

Gene birth and death dynamics are thought to be a primary genomic mechanism underlying functional diversification (Ohno, 1970; Zhang, 2003). A number of aphid-specific gene family expansions have been detected in *A. pisum*, contributing to diverse functions including chemoreception (Smadja *et al*., 2009), development (Shigenobu *et al.,* 2010), and carbon transport (Huerta-Cepas *et al.*, 2010). To characterize the extent of gene duplication in *Ua*, we infer phylogenetic relationships of *Ua* coding sequences with their homologs in other insect species. To do so, we make use of PhylomeDB, a database of gene-based tree topologies, protein alignments and orthology predictions spanning the complete set of coding sequences from dozens of species, including twelve insects (Huerta-Cepas *et al.*, 2014).

The analysis of protein phylogenies represents an approach to comparative genomics that is both gene-centric and genome-wide (Huerta-Cepas *et al*., 2014). It provides a phylogenetic

approach to defining gene families, circumventing the limitations of alternative approaches based on pairwise similarity or genetic distance (Gabaldón, 2008). High-throughput phylogenetic analysis is also an efficient way to detect lineage specific gene expansions (LSEs), which are reflected in tree topologies as monophyletic groups of proteins from one species, relative to a homolog from another species.

Traditionally, duplicated genes are thought to contribute to novel phenotypes primarily via sequence evolution. Evidence exists, however, that sequence divergence and expression divergence are only weakly correlated among duplicate gene pairs, and that in many cases expression divergence is accelerated relative to sequence differentiation (Wagner, 2000). To investigate the role of expression divergence in association with LSE, we use RNA-seq to measure the effect of host plant use on *Ua* expression patterns. These data are also utilized to examine clusters of genes with co-expressed differential expression patterns in response to alternate host plants.

Goals

We sequence the *Ua* transcriptome in pursuit of three major goals. First, we assemble and annotate the feeding transcriptome of *Ua,* providing one of the most extensive genomic resources available for any aphid aside from *A. pisum*. The main motivation is to provide a reference set of transcripts for analysis of intraspecific patterns of genomic differentiation, which we undertake separately (Gill, Chapter 3). Our second goal is gene-based phylogenetic analysis of the full set of coding sequences, to define *Ua* gene families and identify lineage-specific expansions that may underlie host-use or other key functional traits. Third, we assess the role of regulatory evolution using RNA-seq to look for transcripts that are differentially expressed by aphids when feeding on different host plants.

METHODS

Aphid sampling

Ua colonies were collected from five dispersed populations across North America (Table 1). For each population, multiple colonies were sampled from host plants distributed not more than 200 miles apart. Each colony was sampled with minimal disturbance by snipping plant parts on which large aphid colonies were feeding, and placing the plant directly into 25mL Falcon tubes with the aphids still feeding. The tubes were sealed and immediately placed, on-site, into a large liquid nitrogen Dewar flask, transported to Stony Brook University and preserved in a -80° C freezer until processed for sequencing library preparation. All samples were sorted on an aluminum try placed in a dry-ice/ethanol bath (kept at -50-60° C) to separate aphids from plant material and to remove tarsi without risking RNA degradation. Total RNA was extracted from aphid bodies & heads using a standard phenol-chloroform protocol (Trizol), and total RNA quality was verified using a Bioanalyzer (Agilent). All field-sampled RNA extractions used for SNP-based analysis were normalized using duplex-specific thermostable nuclease to preferentially degrade highly abundant transcripts, allowing increased sequence coverage of low-abundance transcripts (RNA-seq libraries, described later, were not normalized). At this stage RNA was pooled by combining equimolar aliquots of RNA extract from all samples derived from a given population, resulting in one pooled RNA sample for each of the five populations. For each pool, poly-adenylated RNA was isolated from total RNA using oligo-D(t) Dynabeads (Invitrogen), polyA-RNA was fragmented by incubation in a zinc ion based buffer (Ambion), and double-stranded cDNA was synthesized using Superscript II reverse transcriptase (Invitrogen). Following end-repair and adenylation, custom paired-end Illumina sequencing adapters (Operon) were ligated to cDNA fragments. Libraries were size selected at 300bp on an agarose gel, and size-selected libraries were amplified using custom-made Illumina paired-end primers. Ampure beads were used for all clean-up steps (Beckman Coulter). Libraries were quantified using qPCR primers specific to Illumina paired-end adapters (Kapa Biosystems). Each library was then prepared for 100nt paired-end sequencing, with distinct barcodes for each population, on an Illumina HiSeq 2000 (Michael Smith Genome Sciences Centre, BC, Canada).

Sequence processing and assembly

Reads were quality filtered using FastQC (Andrews, 2010) and pre-processed with the Fastx-Toolkit (Hannon, 2010). Reads were assembled *de novo* using two assembly programs, Trinity (Haas *et al*., 2013) and ABySS (Simpson *et al.*, 2009). Reads were assembled with ABySS using default settings at every fourth k-mer value from 48 − 92 (after trimming, max read length was 93nt). The resulting 12 assemblies were merged using trans-ABySS, resulting in a *de novo* reference transcriptome assembly. For this assembly, no splice junctions or gene isoforms were inferred because to do so trans-ABySS requires a reference genome, which is lacking for *Ua.* To use Trinity, reads from all pooled libraries were assembled using default settings.

Annotation

Blast2GO was used to annotate assembled contigs (Conesa *et al.,* 2005) by mapping gene ontology (GO) terms to *Ua* unigenes. These assignments were based on querying each unigene to a) the NCBI non-redundant (nr) protein database using BLASTx (e-value < $1x10^{-5}$) (Altschul *et al.*, 1990) and b) the KEGG biological pathway database (Kanehisa *et al.*, 2011). We extracted likely coding sequences from the Trinity assembly using an included script that models all long reading frames to identify a most likely translated sequence. We refer to this set of protein sequences as *U. ambrosiae* reference coding sequences (CDS). All GO enrichment analyses conducted in the study were done by comparing the distribution of GO terms of the test sample in question to the GO distribution for the *Ua* reference sequences, using Fisher's exact test with $p < .05$ to test for significant differences (*i.e.* GO terms with disproportional representation in the test set of sequences), as implemented in Blast2GO (Conesa *et al.,* 2005). To identify *Ua* homologs of host-associated gene families, host-associated pea aphid genes were downloaded from Genbank and used to construct a custom BLAST database. The *Ua* assembly was then queried against the database, and high identity hits (evalue < $1x10^{-5}$) were accepted as putative homologs.

Phylome analysis

Phylomes are catalogs of evolutionary trees of coding sequences from different species. Each tree in a phylome represents the phylogenetic relationships of related proteins (i.e. homologs and paralogs of an ancestral protein) in the species represented. The determination of

which homologs and paralogs are represented in a given tree is based on the parameters used to generate the multiple sequence alignment (MSA) upon which the phylogenetic inference is based. Phylome analysis was conducted by adding *Ua* coding sequences to the existing MSAs that underlie each phylome tree available in the PhylomeDB database. To associate *Ua* sequences with the appropriate MSA, we used USEARCH to identify the most similar *A. pisum* seed sequence for each *Ua* coding sequence (Edgar, 2010). A low identity (ID > 0.6) threshold was used to identify seed sequences since subsequent protein alignment and phylogenetic analysis of *Ua* coding sequences would filter out spurious matches. Using a custom script, the protein alignments associated with each seed sequence were downloaded, and the respective *Ua* sequences were added to each alignment. Multiple sequence alignments of homologous proteins, plus associated *Ua* sequences, were realigned using MUSCLE under default settings (Edgar, 2004) and then trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009). Maximum likelihood phylogenetic analyses were performed with PhyML, using the same parameters as the PhylomeDB tree from which each MSA was derived (Guindon *et al.*, 2005).

Differential expression analysis

A single adult female aptera from an isogenic greenhouse colony (*i.e.* derived from a single female aphid collected in New York) was placed and allowed to reproduce on clippings of each of two replicate plants of each of three species—*Ambrosia trifida* (AT; principal eastern host), *Iva frutescens* (IF; eastern alternate host) and *Tithonia rotundifolia* (TR; southwestern alternate host that is not available to eastern populations). Resulting colonies (female plus all offspring, which ranged in number from 4-9) were collected into vials that were immediately placed in liquid nitrogen to preserve RNA content. Vials were then stored at −80° C until processing. Total RNA was extracted using Trizol (Invitrogen). Poly-A tailed RNA was isolated, and RNA-seq libraries were constructed in-house at Stony Brook University according to a modified RNA-seq protocol (Yoon & Brem, 2010). Each of the six samples was individually barcoded and then all six were pooled. This RNA-seq pool was sequenced in the same lane as the pooled transcriptome libraries described above.

To identify transcripts that are consistently differentially expressed by host plant treatment, we used Trinity's downstream differential analysis pipeline implementing RSEM and edgeR, using default parameters except where noted (Haas *et al*., 2013). Bowtie was used with

default settings to map RNA-seq reads to the *de novo* assembly generated using the pooled population-level data, described above (Langmead *et al.*, 2009). Only forward reads were used, since using both forward and paired reads would have the effect of counting each expressed transcript twice. The following steps were all executed using scripts provided with the Trinity package. RSEM estimates the number of read counts per transcript (i.e. expression level), using likelihood-based posterior probabilities accounting for multiple valid alignments of individual reads to the *Ua* reference assembly (Li & Dewey, 2011). edgeR was then used to estimate whether counts per transcript are significantly different across treatments using a Poisson model to account for biological and technical variability (Robinson *et al.*, 2010). To standardize read counts according to the size of the sequence library generated for each sample as well as the length of each transcript, we performed TMM (trimmed mean of M, an abundance statistic) normalization to obtain FPKM counts (fragments per kilobase of transcript per million) that serve as normalized expression values (Robinson & Oshlack, 2010). To provide a framework for establishing significance cutoffs for differentially expressed transcripts, the distribution of q-values, based on the false discovery rate, was calculated using the software package qvalue (Dabney & Storey, 2014). Default settings were used except that the bootstrap method was used for estimating the null distribution.

RESULTS & DISCUSSION

Sequencing and assembly

Sequencing of all five populations resulted in over 125 million reads totaling 102.8 megabases (Table 2). Coverage by population ranged from a high of 24.3 million reads (AZ) to a low of 16.1 million (AR), with an average sequence coverage of 20.6 million reads per population. Sequence reads were assembled using two software packages, trans-Abyss and Trinity. While the assemblies are of similar size (around 75mb each), the Trinity assembly generated fewer contigs (98,441 compared to nearly 250,000 from trans-Abyss), and the total assembled bases in the Trinity assembly are in contigs 150% longer (n50 = 1436bp) (Table 3). Based on these metrics we retained the Trinity assembly. Short read assembly methods are prone to generate spurious contigs based on sequencing errors. The highly similar sequences that result can confound downstream analysis, since short reads from individual populations may align to different contigs that originate from the same biological transcript. To account for this, we clustered high-identity contigs (ID = 0.9) from the assembly using USEARCH, retaining the longest sequence from each cluster (Edgar, 2010). The resulting filtered, or non-redundant, assembly totals 81,552 transcripts (Table 3) that were used for downstream analysis. We refer to these non-redundant sequences as "isogenes," to distinguish them from the raw contigs generated in the assembly. Open reading frames, or coding sequences (CDS) were inferred in 54.6% (53,760) of the assembled contigs. These are unlikely to all derive from unique proteins, given that the curated pea aphid genome contains only 36,275 CDS, which is likely to be an over-estimate (IAGC, 2010).

To evaluate the overall quality of the assembly we examined the length distribution of *Ua* contigs in comparison to related transcriptome sets (Figure 1). Over half the sequences in the assembly (n = 53,396) are shorter than 400 nucleotides. The proportion of short contigs in the *Ua* assembly is higher than in the pea aphid mRNA set (Legeai *et al.,* 2010) and very few have coding sequences (minimum CDS length is 150bp), suggesting that many of these sequences represent small fragments of longer transcripts. While the majority of longer (> 500bp) unigenes have matches to the NCBI nr database, only about half of the shorter unigenes had any matches.

We compared the *Ua* assembly to mRNA sequences from the pea aphid genome (IAGC, 2010), and to a whitefly transcriptome (Karatolos *et al*., 2011), which represents the most closely

related *de novo* assembly. The *Ua* assembly is similar in total size (74.6 mb) to the pea aphid transcriptome (72.3 mb), but the pea aphid genome has around 37,000 mRNA transcripts of an average length near two thousand bases, compared to 98,441 *Ua* contigs with a mean length of 758nt (Table 3). The assembly contains over 50% more coding sequences than the pea aphid, yet the total size of these coding sequences is little more than half, suggesting that a large number of coding sequences represent fragments. No other high-throughput transcriptome assemblies are available for other aphids (EST sets numbering in the thousands of sequences for several economically important species), leaving the whitefly transcriptome as the closest comparison. The *Ua* assembly compares favorably to the whitefly assembly, with fewer contigs of greater length totaling a similar overall assembly size (Table 3).

To evaluate the proportion of assembled contigs that approach full-length records of expressed transcripts, we calculated the ortholog hit ratio (OHR), the length ratio of each *Ua* sequence in relation to its best BLAST match in the pea aphid. While the assembly contains many fragmentary contigs (OHR << 1) that are quite short (< 400bp), a significant proportion cluster around OHR=1, indicating sequences at or near full length (Figure 2). The intactness of these longer transcripts corresponds to the pattern in Figure 1, which shows large mismatches between numbers of *Ua* transcripts, *Ua* CDS sequences and *Ap* mRNA sequences at low sequence lengths, but agreement for longer sequences. We conclude that although these data do not represent a complete record of the expressed genes in *U. ambrosiae*, they do cover a large number of complete transcripts and represent a more complete transcriptome than is available for any aphid species aside from *A. pisum*.

Annotation

Transcripts were annotated using Blast2GO, based on comparison of each sequence to the NCBI database of non-redundant proteins (nr). Over half (52,027 = 52.9%) of *Ua* contigs had at least one match to a known protein, consistent with the expectation that a considerable portion of the assembly should comprise relatively conserved orthologs of proteins known in other branches of life. 65% of these matches were most similar to a pea aphid protein (Figure 3). Sequence-based signatures of functional protein domains within *Ua* contigs were obtained by searching the InterProScan database, resulting in 11,050 matches. We used Blast2GO to map GO terms to 30,647 *Ua* contigs, almost a third of the total assembly. Contigs not assigned GO terms

could be derived from non-translated transcripts, un-alignable fragments, or novel proteins specific to the *Uroleucon* lineage or its ancestors since they split from the common ancestor with the pea aphid. Open reading frames (ORFs) were identified in 54.6% (53,760) of the assembled contigs. The resulting coding sequences were compiled into a *Ua* protein sequence set.

Host-associated genes

We used reciprocal best-hit BLAST matches to identify *Ua* orthologs of known host-associated genes in the pea aphid. In total, 419 *Ua* contigs were assigned as orthologs to host-use associated pea aphid gene families. These include chemosensory proteins and receptors (29 *Ua* sequences matched a database of 105 pea aphid genes, a match rate of 27.6%), p450 detoxification enzymes (50/85; 58.9%), salivary secretion constituents (163/279; 58.4%), circadian clock genes (37/13; 280%), and ion channel genes (140/93; 151%). Since we did not sequence and assemble the entire genome, or even the entire set of *Ua* expressed sequences, we cannot conclude that we've identified the full set of *Ua* homologs of host-associated gene families. Nevertheless, the homologs identified provide a sufficient resource to characterize patterns of variation in host-associated genes compared to loci without evident functional relevance to host use. Despite the incompleteness of this transcriptome set we recover greater numbers of circadian and ion channel genes than are present in *A. pisum*, suggesting a possible expansion of these gene families in *Uroleucon* species. This possibility is investigated in the phylome analysis discussed below.

Differential expression analysis

In order to gauge whether regulatory evolution may be an element in divergent host use in *Ua,* we subject aphids of a single genotype to three host plant treatments—*Ambrosia trifida* (AT), *Iva frutescens* (IF), and *Tithonia rotundifolia* (TR)—with two replicates per treatment, and used RNA-seq to quantify differential expression. Sequencing of the six RNA-seq sub-libraries resulted in nearly 240 million sequence reads (Table 2). These reads were aligned to all 98,441 contigs in the assembly. Using RSEM's probabilistic model that accounts for multiple valid alignments, 18,852 contigs had minimum coverage of at least ten reads from each sample.

The estimated number of read counts per transcript derived from these alignments was used to identify differentially expressed transcripts, defined as showing a fold change factor of at

least two among host plant treatments. Determining appropriate significance cutoffs for RNA-seq experiments can be problematic due to the large number of features (*i.e.* transcripts) tested, which can result in a large number of false positives at commonly used p-value significance cutoffs (Storey, 2003). It is helpful to explicitly estimate the number of false positives expected for given p-value cutoffs, which can be achieved using the distribution of q-values based on the false discovery rates associated with each p-value. The q-value distribution is based on estimating the "true" proportion of null features (π_o), *i.e.* non-differentially expressed transcripts, given the p-value distribution. We find $\pi_o = 0.96$. Thus we expect 4% of the tested transcripts to be differentially expressed from this experiment. This equates to an expectation that 754 transcripts are differentially expressed (4% of 18,852 transcripts in the analysis), and provides a basis to estimate the proportion of false positives expected given a specific p-value cutoff (Storey, 2002). We elected to accept a significance cutoff of $p<0.0005$. Although this may be considered an especially stringent cutoff, the q-value distribution indicates that more relaxed cutoffs entail increasing rates of false discovery (Figure 4). This cutoff yields 112 differentially expressed transcripts with a q-value of 0.054, corresponding to an expectation that 5.4% (6 transcripts) will be false positives.

We next assessed the degree of correlation in expression profiles among the 112 differentially expressed transcripts by generating a distance tree based on hierarchical clustering of correlation values (Figure 5). As expected, comparative analysis of expression correlations indicates that AT replicates are highly similar (forming a sister group in the tree at the top of Figure 5); the same is true for IF replicates. The TR replicates appear to be inconsistent with one another, as suggested by their paraphyletic placement in the distance tree at the top of Figure 5, where many transcripts are anti-correlated in the two TR treatments—up-regulated (green) in one replicate and down-regulated (red) in the other. We therefore exclude TR treatments from downstream analysis. This leaves 91 transcripts that are significantly differentially expressed with minimum fold change of 2 in IF vs. AT treatments, at a false discovery rate of $q=0.54$.

Two GO terms—protein kinase activity and cellular amino acid metabolic process—were enriched ($p < 0.05$) among differentially expressed transcripts, each represented by four contigs. Protein kinases are biologically active components of plant phloem, many of which interact with calcium ions (Will, 2006). Aphid effectors are known to inhibit plant defensive responses by manipulating calcium-mediated signaling pathways (Hogenhout & Bos, 2011). This

points to the possibility that differential expression of protein kinases is driven by differences in plant phloem chemistry. The other over-represented GO category is cellular amino acid metabolism. The well-described co-dependence of aphids with their *Buchnera* endosymbionts is mediated by interrelated amino acid pathways (Hansen & Moran, 2011), a benefit to aphids given their relatively simple, protein-poor diet of plant sap. In general, amino acid synthesis is a challenge for aphids, and expression levels of genes related to these pathways may respond to species-specific phloem profiles in their hosts.

In addition to analysis of differential expression on a gene-by-gene basis, clusters of transcripts with correlated differential expression profiles in response to host plant treatment were identified. Each of these clusters represents a group of *Ua* transcripts with highly correlated expression profiles that are divergent on AT compared to IF, *e.g.* each cluster member is over-expressed in both AT replicates, and down-regulated in both IF replicates (Figure 6). For the purpose of identifying clusters, the expression results for IF were ignored due to considerable variance among the two IF replicates, which can be seen in the visualized cluster co-expression patterns (Figure 6b), as well as in Figure 5.

Fourteen clusters totaling 91 isogenes were identified as being differentially co-expressed on AT versus IF. Of these, 9 are assigned at least one GO term related to purine metabolic processes. Five are grouped in a single cluster (Figure 6a) while four others are each in separate clusters. Purine metabolism in aphids is closely coupled with *Buchnera* metabolism, with key genes in the pathway absent in one species but present in the other (Ramsey *et al.* 2010). Purine metabolism is one co-dependency among a network of amino acid processing functions shared between the two species (Hansen & Moran, 2011). Several of the genes associated with purine functioning are also associated with GO terms for adenine metabolism and nucleoside binding, two other processes that physiologically link aphids and *Buchnera*. These results add to evidence that the aphid-*Buchnera* symbiosis responds dynamically to different host plants and may be intimately involved in aphid host-use evolution (Oliver *et al.,* 2010). Other GO functions are not as suggestive of a direct role in host use. For example, several genes relate to Kappa B-kinase activity, which is implicated in stress response, especially in relation to changing bacterial abundance (Riddell *et al.,* 2011)

Phylogenetic analysis of coding sequences

We took a phylogenetic approach to inferring orthologous relationships between *Ua* and *Ap* proteins by merging *Ua* coding sequences to the multiple sequence alignments of homologous proteins available in the PhylomeDB database, which includes the complete pea aphid gene set (Huerta-Cepas *et al*., 2014). We first identified 34,210 *Ua* coding sequences (63.6% of all CDS) with genetic distance-based similarity to pea aphid genes. Each pea aphid gene serves as a "seed" sequence for its respective gene tree, which provides phylogenetic relationships based on a multiple sequence alignment (MSA) of homologous genes in the pea aphid and eleven other insect species.

31,112 unique *Ua* coding sequences were successfully aligned to an existing PhylomeDB MSA. As expected given the genealogical relatedness of similar genes, in a number of cases multiple *Ua* CDS aligned to the same MSA, resulting in 8,802 total gene trees. Within these trees, *Ua* CDS grouped into 17,826 unique "*Ua*-only" groups, defined as the largest clades containing only *Ua* sequences that are sister to a non-*Uroleucon* species.

The main goal with phylome analysis is to characterize the size and number of *Ua*-only clades with one-to-one, few-to-one, and many-to-one relationships with non-*Ua* genes, centrally pea aphid genes. One-to-one relationships describe singleton *Ua* coding sequences with sister relationships to non-*Ua* genes. We arbitrarily use a cutoff of 10 *Ua* sequences to delineate few-to-one and many-to-one groupings, such that the former describe *Ua*-only clades of two to nine *Ua* coding sequences, and many-to-one relationships are *Ua*-only clades of ten or more sequences that are sister to a single non-*Uroleucon* CDS. Examples of trees with different clade sizes are illustrated in Figure 7.

One pitfall of a fragmentary transcriptome is the potential for some of the large number of short CDS to contribute to mis-alignments. Non-overlapping fragments of individual *Ua* sequences could inflate the number of few-to-one or many-to-one groups by representing multiple leaves in a clade, despite originating from a single transcript. Alternatively, fragments with a common origin may each align to distinct MSAs, leading to spurious one-to-one groups. We address the effect of short transcript sequences on *Ua*-only group size by first examining correlations between group size and mean group sequence length, and second by plotting the distribution of ortholog hit ratio values among the three size classes.

Two features are of note in the correlation of group sizes and mean group sequence lengths (Figure 8). First, the correlation coefficients (all below R^2=0.085) are far closer to 0 than 1,

indicating only a small relative effect of sequence length on group size. The effect of sequence length in many-to-one groups (R^2=0.053) is not significantly different than that for the total set of *Ua*-only groups (R^2=0.0824), based on Fisher's z-transformation to test for differences of correlation coefficients (p=0.18). Second, groups with short mean lengths are largely present in one-to-one and few-to-one clades—the bottom left portion of the plot. There is little evidence that a significant number of many-to-one groups are comprised of fragmentary sequences, helping to justify a focus on many-to-one groups of at least ten *Ua* transcripts in our assessment of lineage-specific expansion events.

The distribution of ortholog hit ratio (OHR) values for *Ua*-only groups suggests that the major concentration of CDS are at or near full length, with OHR > 0.75, for all three group sizes (Figure 9). To minimize the presence of fragmentary sequences, we exclude all sequences with OHR<0.75 from further analysis. Low OHR values are excluded, rather than high OHR values retained, to avoid discarding sequences without OHR values, which may represent novel or divergent genes lacking the high similarity to any *A. pisum* coding sequences that is required to calculate the OHR. This cutoff excludes 1386 one-to-one *Ua* sequences (21.9%), 2235 few-to-one (11.3%), and 174 (3.5%) of many-to-one sequences.

In total, *Ua* CDS are represented in 17,826 *Ua*-only clades ranging in size from 1 to 51 coding sequences. Figure 10 shows the frequency distribution of all *Ua*-only clades of various sizes inferred in the analysis. We focus on the right tail of this distribution, minus sequences with OHR < 0.75. This yields 4758 transcripts in many-to-one clades, or 8% of the 58,604 *Ua* CDS represented in one of the 8802 phylome trees in the analysis. 4950 CDS had one-to-one relationships with pea aphid or other non-*Uroleucon* genes, and 17,609 CDS are in *Ua*-only groups that are sister to an *A. pisum* gene. For *A. pisum,* 4059 protein sequences had one-to-one relationships with homologs in other species, and 2282 had many-to-one relationships (Huerta-Cepas *et al*., 2010). Thus *Ua* shows similar rates of gene expansion as seen in the pea aphid (Table 4).

Lineage-specific gene expansions (LSE) are a major process underlying coding and regulatory diversity in eukaryotes, and are thought to be one of the principal means of adaptation (Lespinet, 2002). We define LSE as the proliferation of paralogous genes via successive duplication events in one lineage relative to a second—in this case, *U. ambrosiae* relative to *A. pisum*. The presence of many-to-one relationships represents likely gene family expansions

based on successive duplication events that followed the divergence of *Ua*'s common ancestor with that of the pea aphid. LSEs generate paralogs, which originate as duplicated genes and often evolve to acquire novel functions.

We have limited ability to determine whether *Ua* gene families have undergone gene loss, since the incompleteness of the assembly means we cannot assume that "missing" genes represent true gene losses rather than missing data. However, 4950 *Ua* coding sequences had one-to-one relationships with *Ap* coding sequences, and 299 of these *Ua* sequences appear in multiple clades. Each of these 299 sequences can thus be phylogenetically associated with multiple *Ap* proteins, suggesting the corresponding *Ap* families underwent lineage-specific gene expansion, or *Ua* has experienced lineage-specific gene loss.

GO enrichment analysis indicates that 89 terms are over-represented among LSE gene families (FDR < 0.005) (Table 5). Of particular interest are functional categories for which gene expansion has previously been shown to be associated with functional traits related to host use. The most striking example of LSE we detect is for serine-type endopeptidase inhibitor activity, assigned to 62 *Ua* coding sequences. Serine-type endopeptidase is a proteinase inhibitor (PI) mobilized by plants into the phloem (Broadway, 1997). PIs have deleterious effects on some herbivorous insects, though many other insect species, including many aphids, have evolved counter-adaptations (Jongsma & Bolter, 1997). They may also play a defensive role in inhibiting digestive enzymes secreted by insects into the phloem. Intriguingly, insect resistance to PIs is host-specific, and plays a potential role in delineating host breadth—the generalist caterpillar *Spodoptera frugiperda* has evolved a biochemical mechanism to alter endopeptidase enzymes, whereas specialist herbivores tend not to have a wide enough spectrum of modification enzymes to overcome non-host PIs (Falco & Silva-Filho, 2003). Expansion of this gene family in *Uroleucon* may provide a reservoir of enzymes available to counteract the defensive responses of plants with which they do not have a long co-evolutionary history. This possibility leads to the testable prediction that generalist species might retain more expansive host-associated gene families than specialists. This is the pattern observed in the specialist *Drosophila sechellia*, which has undergone extensive and rapid gene loss compared to its more generalist relative *D. simulans* (McBride, 2007).

In comparing many-to-one *Ua* gene families to the set of *a priori* host associated families identified above, we find additional evidence that LSE may play a direct role in host-use

evolution. Thirteen salivary gland genes are represented among the expanded gene families we identify, as are 23 circadian genes. Aphids secrete salivary proteins beginning with the initiation of probing. Salivary secretions include digestive proteins, proteins that deal with the physical challenges of sap-feeding (*e.g.* maintaining osmotic pressure), and proteins that interact as effectors with plant defense and immune systems (Carolan *et al*., 2011; Will *et al.*, 2007). Gene family expansion of salivary proteins may allow paralogs to adaptively evolve in response to novel phloem profiles without losing ancestral functions.

Circadian clock genes also appear to have undergone expansion. The molecular basis of circadian clocks is highly conserved across eukaryotes, making it somewhat unexpected to find that circadian clock genes exhibit accelerated rates of change in *A. pisum* relative to *D. melanogaster* (Cortés *et al*., 2010). It remains unclear what selective forces shape rapid evolution of circadian clock genes. Regardless of their uncertain function, lineage-specific expansion of these genes in *Uroleucon* adds to evidence that circadian gene families continued to diversify as aphid lineages have diverged.

Expression patterns in expanding gene families

It has long been appreciated that gene duplication contributes to evolutionary novelties through neo- or sub-functionalization of duplicated genes (Ohno, 1970). More recently, it has been proposed that duplicated genes may also undergo rapid expression evolution, resulting in divergent expression profiles of closely related genes (Wagner, 2000). In particular, the pattern observed in yeast is that following duplication, expression evolution is rapid for one copy while the other tends to retain ancestral expression profiles, resulting in expression asymmetry (Gu *et al.,* 2005). To test the hypothesis that gene family expansion is accompanied by divergent expression patterns among paralogs, we compared the variance of fold change values for transcripts in many-to-one, few-to-one, and one-to-one *Ua*-only groups. The variance of putatively duplicated coding sequences (*i.e.* CDS in many-to-one and few-to-one groups) was nearly identical (37.68 and 37.73, respectively). Many-to-one transcripts have a significantly higher fold change variance than transcripts with one-to-one homology relationships (30.82; F-test, $p = .018$). Few-to-one transcripts are also significantly higher ($p = .01$). These results tentatively support the view that paralogous genes tend to evolve divergent expression profiles.

Expression variances were also calculated for *a priori* host-use associated gene families.

The variance for all such genes (37.69) is greater than that of the transcriptome-wide variance (30.74), but the difference is not significant (p=0.3). The lack of significance could be an artifact of the very small sample size of *a prior* gene families (n=37 for those sequences that meet differential expression cutoffs) compared to the assembly as a whole. However, the large variance of host-use genes is more likely due to large differences among, rather than within, particular classes of host-use genes. Variances for these genes range from a low of 3.16 for chemosensory genes to a high of 29.8 for ion channel genes.

Conclusion

The annotated transcriptome reported here represents a major enhancement of the genetic resources available for the study of intraspecific evolution in *U. ambrosiae*, for which only a handful of sequenced genes were previously available. We identify *a priori* host-associated gene families to guide further analysis of host-use divergence among *Ua* populations. We find that lineage-specific gene expansions are an important element of genome evolution in *Ua*, including for loci that directly interact with plant hosts. Expansions of key gene families may help determine host-breadth, among other traits, by providing diverse genes capable of responding to diverse host plant conditions. Although it is difficult to make any conclusive statements based on our differential expression analysis, since only one genotype on two host plants was examined, we do find that gene expansion is associated with expression variation. A worthy goal for future research will be to assess the relative contributions of gene diversification, expression evolution and sequence divergence in shaping host use evolution. *U. ambrosiae* represents a promising system to pursue this line of research.

Tables

Table 1. Source of aphid samples used for transcriptome sequencing. Aphids were collected from multiple localities in each of four populations. All samples were taken from distinct host plant individuals. Eastern specialist populations were collected from *Iva frutescens* while generalist populations were collected from a variety of host plants.

Sample ID	Sampling localities	No. colonies sampled	Host breadth	Host taxa
NE (northeast)	NY, MA, NH	10	Specialist	I. frutescens
AC (mid-Atlantic Coast)	VA, NJ	5	Specialist	I. frutescens
OH (Ohio)	OH	4	Specialist	A. trifida
AR (Arkansas)	AR	6	Generalist	Ambrosia, Bidens, Eupatorium
SW (southwest)	AZ, NM	15	Generalist	Ambrosia, Viguiera, Tithonia, Heterotheca

Table 2. Results in numbers of raw reads from sequencing twelve barcoded sub-libraries on a single lane on an Illumina Hi-Seq 2000. The top 5 samples are based on field-collected aphids and represent the pooled population samples. The bottom six samples comprise two replicates of a single eastern aphid genotype for each of three host plant treatments. AT = *Ambrosia trifida,* TR = *Tithonia rotundifolia,* IF = *Iva frutescens*.

	For. paired	Rev. paired	For. unpaired	Rev. unpaired	TOTAL
Pooled samples					
NE	6.77	6.98	3.51	0.97	18.23
AC	8.03	7.42	5.46	0.84	21.74
AR	6.94	6.31	2.06	0.74	16.06
OH	8.77	8.91	3.68	1.09	22.45
AZ	10.15	9.43	3.63	1.12	24.34
					102.81
RNAseq samples					
AT-1	18.75	17.12	4.39	1.85	42.12
AT-2	18.20	16.48	6.16	3.73	44.57
TR-1	18.20	16.76	4.25	2.12	41.34
TR-2	14.53	13.06	3.00	1.64	32.24
IF-1	17.90	16.58	4.06	1.75	40.29
IF-2	17.12	16.00	4.17	1.67	38.97
					239.52

Table 3. Results for all sequence sets discussed in the study. These include the *U. ambrosiae* Trans-ABySS assembly, and the Trinity assembly both in its entirety and clustered at ID=0.9— the final *Ua* isogene reference used for downstream analysis. Inferred open reading frames, or coding sequences (CDS), are also presented. Also given are mRNA sequences from the *A. pisum* whole genome assembly (IAGC, 2010) and transcriptome assemblies from the greenhouse whitefly *Trialeurodes vaporariorum* (Karatolos *et al.,* 2011). All values are in units of basepairs unless otherwise noted. N50: over half the assembled bases are in contigs of this length or longer; max = maximum contig length; sum = total length of assembly in megabases.

Sequence set	no. contigs	N50	max	mean	sum (Mb)	Assembly method	Sequencing tech.
Ua assembly (Trans-ABySS)	249689	938	10,676	382	77	Trans-ABySS	Illumina
Ua assembly (Trinity)	98,441	1,436	15,734	758	74.6	Trinity	Illumina
Ua reference transcripts	81,552	809	15,734	571	46.6	Trinity (filtered)	Illumina
Ua open reading frames (CDS)	53,760	292	982	741	3.90		
A. pisum mRNA	36,961	2,628	62,264	1,956	72.30	from genome assembly	Illumina
T. vaporariorum transcriptome (Velvet)	253,603	--	6,350	312	79.12	Velvet	Illumina
T. vaporariorum (est2assembly)	54,748	--	--	965	5.28	est2assembly	454

Table 4. Results from phylome analysis. The first column lists the three size classes of *Ua*-only clades. Column 2 gives the total number of phylome trees that include at least one *Ua* coding sequence. Column 3 lists the total number of monophyletic groups of *Ua* CDS that are sister to a non-*Uroleucon* sequence. A given phylome tree may have multiple *Ua*-only clades. Column 4 gives the number of unique *Ua* coding sequences represented in each size class. The final column lists the number of unique *Ua* CDS in each size class after excluding those with ortholog hit ratio of OHR<0.75.

Ua-only clade size group	No. trees	No. clades	No. unique CDS	No. with high OHR
one-to-one	5674	6635	6336	4950
few-to-one	5572	10083	19844	17609
many-to-one	346	952	4932	4758

Table 5. Enriched GO terms (false discovery rate = FDR < .05) for coding sequences that are in *Ua*-only clades of at least ten sequences. Enrichment refers to an over-abundance of GO terms in a test set, in this case *Ua* sequences in many-to-one clades, compared to the total set of annotated *Ua* contigs. Each GO term is categorized as a molecular function (F), cellular component (C), or biological process (P).

GO-ID	Category	FDR	No. CDS	GO term
GO:0004867	F	2.02E-36	62	serine-type endopeptidase inhibitor activity
GO:0042302	F	7.34E-18	54	structural constituent of cuticle
GO:0005524	F	5.47E-16	366	ATP binding
GO:0009343	C	2.48E-11	25	biotin carboxylase complex
GO:0004075	F	4.83E-11	25	biotin carboxylase activity
GO:0016199	P	5.63E-11	13	axon midline choice point recognition
GO:0003878	F	3.33E-10	21	ATP citrate synthase activity
GO:0031071	F	3.96E-10	15	cysteine desulfurase activity
GO:0004775	F	1.48E-09	21	succinate-CoA ligase (ADP-forming) activity
GO:0042709	C	2.88E-09	21	succinate-CoA ligase complex
GO:0003746	F	3.04E-09	38	translation elongation factor activity
GO:0007527	P	3.31E-09	11	adult somatic muscle development
GO:0042073	P	3.67E-09	13	intraflagellar transport
GO:0017069	F	4.30E-09	12	snRNA binding
GO:0045793	P	4.30E-09	12	positive regulation of cell size
GO:0045214	P	2.43E-08	12	sarcomere organization
GO:0008097	F	2.43E-08	12	5S rRNA binding
GO:0038007	P	2.48E-08	10	netrin-activated signaling pathway
GO:0070593	P	2.87E-08	14	dendrite self-avoidance
GO:0017137	F	2.99E-08	11	Rab GTPase binding
GO:0000398	P	2.99E-08	25	mRNA splicing, via spliceosome
GO:0015935	C	1.33E-07	30	small ribosomal subunit
GO:0004347	F	1.56E-07	11	glucose-6-phosphate isomerase activity
GO:0021551	P	1.56E-07	13	central nervous system morphogenesis
GO:0048846	P	1.56E-07	13	axon extension involved in axon guidance
GO:0008046	F	1.56E-07	13	axon guidance receptor activity
GO:0019992	F	1.74E-07	9	diacylglycerol binding
GO:0005219	F	1.74E-07	9	ryanodine-sensitive calcium-release channel activity
GO:0006378	P	1.94E-07	10	mRNA polyadenylation
GO:0008518	F	1.94E-07	10	reduced folate carrier activity
GO:0005849	C	1.94E-07	10	mRNA cleavage factor complex
GO:0006094	P	3.16E-07	46	gluconeogenesis
GO:0032851	P	3.50E-07	16	positive regulation of Rab GTPase activity
GO:0005932	C	3.98E-07	13	microtubule basal body
GO:0007413	P	3.98E-07	13	axonal fasciculation
GO:0016459	C	4.61E-07	31	myosin complex
GO:0004174	F	5.48E-07	11	electron-transferring-flavoprotein dehydrogenase activity
GO:0005097	F	6.49E-07	16	Rab GTPase activator activity
GO:0004634	F	8.18E-07	12	phosphopyruvate hydratase activity
GO:0000015	C	8.18E-07	12	phosphopyruvate hydratase complex

CHAPTER 3 REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Berenbaum, M. (2002). Postgenomic chemical ecology: from genetic code to ecological interactions. Journal of Chemical Ecology, 28(5), 873–896.

Bernays, E. A., & Chapman, R. F. (1994). Host-plant selection by phytophagous insects. Springer.

Brisson, J. A., & Stern, D. L. (2006). The pea aphid, *Acyrthosiphon pisum*: an emerging genomic model system for ecological, developmental and evolutionary studies. BioEssays, 28(7), 747–755.

Broadway, R. M. (1997). Dietary regulation of serine proteinases that are resistant to serine proteinase inhibitors. Journal of Insect Physiology, 43(9), 855–874.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973.

Carolan, J. C., Caragea, D., Reardon, K. T., Mutti, N. S., Dittmer, N., Pappan, K. (2011). Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): A dual transcriptomic/proteomic approach. Journal of Proteome Research, 10(4), 1505–1518.

Carolan, J. C., Fitzroy, C. I. J., Ashton, P. D., Douglas, A. E., & Wilkinson, T. L. (2009). The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. Proteomics, 9(9), 2457–2467.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674–3676.

Cortés, T., Ortiz-Rivas, B., & Martínez-Torres, D. (2010). Identification and characterization of circadian clock genes in the pea aphid *Acyrthosiphon pisum*. Insect Molecular Biology, 19, 123–139.

Dabney A and Storey JD. qvalue: Q-value estimation for false discovery rate control. R package version 1.38.0.

Dworkin, I., & Jones, C. D. (2008). Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics, 181(2), 721–736.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460–2461.

Falco, M. C., & Silva-Filho, M. C. (2003). Expression of soybean proteinase inhibitors in transgenic sugarcane plants: effects on natural defense against *Diatraea saccharalis*. Plant Physiology and Biochemistry, 41(8), 761–766.

Ferrari, J., Via, S., & Godfray, H. C. J. (2008). Population differentiation and genetic variation in performance on eight hosts in the pea aphid complex. Evolution; International Journal of Organic Evolution, 62(10), 2508–2524.

Frantz, A., Plantegenest, M., Mieuzet, L., & Simon, J.C. (2006). Ecological specialization correlates with genotypic differentiation in sympatric host-populations of the pea aphid. Journal of Evolutionary Biology, 19(2), 392–401.

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biology, 9(10), 235–235.

Giordanengo, P., Brunissen, L., Rusterucci, C., Vincent, C., Van Bel, A., Dinant, S. (2010). Compatible plant-aphid interactions: How aphids manipulate plant responses. Comptes Rendus Biologies, 333(6-7), 516–523.

Gu, X., Zhang, Z., & Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proceedings of the National Academy of Sciences, 102(3), 707–712.

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research, 33(Web Server issue), W557–W559.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494–1512.

Hansen, A., & Moran, N. (2011). Aphid genome expression reveals host–symbiont cooperation in the production of amino acids. Proceedings of the National Academy of Sciences, 108(7), 2849-2854.

Hawthorne, D., & Via, S. (2001). Genetic linkage of ecological specialization and reproductive isolation in pea aphids. Nature, 412(6850), 904–907.

Hogenhout, S. A., Van der Hoorn, R. A. L., Terauchi, R., & Kamoun, S. (2009). Emerging concepts in effector biology of plant-associated organisms. Molecular Plant-Microbe Interactions, 22(2), 115–122.

Hogenhout, S. A., & Bos, J. I. (2011). Effector proteins that modulate plant–insect interactions. Current Opinion in Plant Biology, 14(4), 422–428.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., & Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Research, 42(Database issue), D897–902.

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

International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biology, *8*(2), e1000313, 1-24.

Jean, P., & Jean-Christophe, S. (2010). The pea aphid complex as a model of ecological speciation. Ecological Entomology, 35, 119–130.

Jongsma, M. A., & Bolter, C. (1997). The adaptation of insects to plant protease inhibitors. Journal of Insect Physiology, *43*(10), 885–895.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research, 40(D1), D109–D114.

Karatolos, N., Pauchet, Y., Wilkinson, P., Chauhan, R., Denholm, I., Gorman, K. (2011). Pyrosequencing the transcriptome of the greenhouse whitefly*, Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. BMC Genomics, 12(56), 1-14.

Kim, H., Lee, S., & Jang, Y. (2011). Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): Diversification, host association, and biogeographic origins. PLoS One, 6(9), e24749, 1-17.

Kopp, A., Barmina, O., Hamilton, A. M., Higgins, L., Mcintyre, L. M., & Jones, C. D. (2008). Evolution of gene expression in the *Drosophila* olfactory system. Molecular Biology and Evolution, 25(6), 1081–1092.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25.1–R25.10.

Legeai, F., Shigenobu, S., Gauthier, J. P., Colbourne, J., Rispe, C., Collin, O. (2010). AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Molecular Biology, 19, 5–12.

Lespinet, O. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Research, 12(7), 1048–1059.

Li, B., & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(323), 1-16.

McBride, C. (2007). Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences, 104(12), 4996-5001.

Mutti, N. S., Louis, J., Pappan, L. K., Pappan, K., Begum, K., Chen, M.-S. (2008). A protein from the salivary glands of the pea aphid, *Acyrthosiphon pisum*, is essential in feeding on a host plant. Proceedings of the National Academy of Sciences, 105(29), 9965–9969.

Ohno, S. (1970). Evolution by gene duplication. New York: Springer-Verlag.

Oliver, K. M., Degnan, P. H., Burke, G. R., & Moran, N. A. (2010). Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. Annual Review of Entomology, *55*, 247–266.

Ollivier, M., Legeai, F., & Rispe, C. (2010). Comparative analysis of the *Acyrthosiphon pisum* genome and expressed sequence tag-based gene sets from other aphid species. Insect Molecular Biology, 19, 33–45.

Ramsey, J. S., Macdonald, S. J., Jander, G., Nakabachi, A., Thomas, G. H., & Douglas, A. E. (2010). Genomic evidence for complementary purine metabolism in the pea aphid,

Acyrthosiphon pisum, and its symbiotic bacterium *Buchnera aphidicola*. Insect Molecular Biology, *19*, 241–248.

Ramsey, J., Rider, D., Walsh, T., De Vos, M., Gordon, K., Ponnala, L. (2010). Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae.* Insect Molecular Biology, 19, 155–164.

Riddell, C. E., Sumner, S., Adams, S., & Mallon, E. B. (2011). Pathways to immunity: temporal dynamics of the bumblebee (*Bombus terrestris*) immune response against a trypanosomal gut parasite. Insect Molecular Biology, *20*(4), 529–540.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), R25.1 – R25.10.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140.

Shigenobu, S., Bickel, R. D., Brisson, J. A., Butts, T., Chang, C. C., Christiaens, O. (2010). Comprehensive survey of developmental genes in the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes. Insect Molecular Biology, 19, 47–62.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19(6), 1117–1123.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J., & Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution, 66(9), 2723–2738.

Smadja, C., Shi, P., Butlin, R. K., & Robertson, H. M. (2009). Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. Molecular Biology and Evolution, 26(9), 2073–2086.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 479–498.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Science, 100, 9440-9445.

Thompson, G. A. (2006). Transcriptomics and functional genomics of plant defence induction by phloem-feeding insects. Journal of Experimental Botany, 57(4), 755–766.

Via, S., & Hawthorne, D. (2002). The genetic architecture of ecological specialization: correlated gene effects on host use and habitat choice in pea aphids. American Naturalist, 159(3), 76–88.

von Dohlen, C., Rowe, C., & Heie, O. (2006). A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. Molecular Phylogenetics and Evolution, 38(2), 316–329.

Vieira, F., & Rozas, J. (2011). Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the

chemosensory system. Genome Biology and Evolution, 3, 476-490.

Wagner, A. (2000). Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. Proceedings of the National Academy of Sciences, 97(12), 6579–6584.

Will, T. (2006). Physical and chemical interactions between aphids and plants. Journal of Experimental Botany, 57(4), 729–737.

Will, T., Tjallingii, W. F., Thönnessen, A., & van Bel, A. J. E. (2007). Molecular sabotage of plant defense by aphid saliva. Proceedings of the National Academy of Sciences of the United States of America, 104(25), 10536–10541.

Yoon, O. K., & Brem, R. B. (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA, 1–13.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6), 292–298.

Chapter 4.

Functional analysis of population-level transcriptome sequences reveals candidate genes for divergent host-use evolution in a non-model aphid species, *Uroleucon ambrosiae*

INTRODUCTION

Biologists over the past several decades have described several macroevolutionary patterns that help explain the evolutionary forces responsible for the unparalleled success of insects. These patterns broadly support the premise popularized by Ehrlich and Raven (1964) that diversification in herbivorous insect lineages is intimately associated with the evolution of host plant use. For example, a large proportion of speciation events are accompanied by host shifts (Winkler & Mitter, 2008), which are often followed by radiations (Futuyma & Agrawal, 2009). Most insect herbivores specialize on closely related host plant groups (Schoonhoven *et al.,* 2005), although transitions are common between generalized and specialized host-breadth through time (Nosil, 2002). This could suggest that lineages diversify via repeated cycles of acquisition of new hosts followed by radiation through host-specialization (Janz & Nylin, 2008). Related species commonly experience host-associated ecological divergence (Shafer & Wolf, 2013; Winkler & Mitter, 2008), which is positively correlated with reproductive isolation across disparate lineages (Funk *et al*., 2006).

These macroevolutionary patterns together suggest that divergent ecological selection formed by local differences in host use may encourage the evolution of reproductive barriers and eventually new species. With the rise of genome-scale sequencing and analysis approaches in non-model systems, research is increasingly taking a microevolutionary focus on the population genomics of ecological divergence (Nosil & Feder, 2012; Nosil & Schluter, 2011). If divergent host-associated selection is a plausible and consistent cause, rather than consequence, of speciation, populations must maintain differentiation at host-associated loci (*i.e.* loci under selection) despite ongoing gene flow (Nosil & Feder, 2012). Yet most genomics work in plant-insect evolution targets species-level diversification, concerning taxa that are already partially if not fully reproductively isolated. As a result, the genetic changes reflecting divergent ecological adaptation (*e.g.* the buildup of genomic differentiation due to divergent selection at ecologically

responsive loci) may be obscured by evolutionary processes, whether neutral or functional, following reproductive isolation.

Here, we characterize genomic differentiation in a non-model system, the brown ambrosia aphid, *Uroleucon ambrosiae* (*Ua*), in order to test for divergent host-associated ecological selection. Specifically, we aim to a) measure allelic differentiation across expressed transcripts, and b) identify candidate functional loci responsible for divergent host-use traits (i.e. generalist versus specialist host breadth) among *Ua* populations. To accomplish this we sequenced the transcriptome (*i.e.* expressed poly-adenylated mRNA transcripts) from five *Ua* populations to identify genome-wide single nucleotide polymorphisms (SNPs). Sampled populations are geographically isolated and exhibit either specialist or generalist host-use phenotypes, based on collecting records and field observations. This multiple-comparison sampling design allows us to isolate the effects of geographic isolation versus host-use divergence on patterns of genetic differentiation among *Ua* populations.

The Uroleucon *study system*

U. ambrosiae has merit as a study system primarily because its populations demonstrate well-described divergence in host-use traits. In the eastern portion of its North American range, populations feed and reproduce on Great Ragweed (*Ambrosia trifida)*, and the closely related *Iva frutescens* in coastal salt marsh habitats, avoiding other co-distributed Asteraceae (Compositae) species (Funk & Bernays, 2001; Moran, 1984; Robinson, 1985). Southwestern populations have a more generalist host-breadth: colonies can be found roughly as often on *A. trifida* as on species in over a dozen other genera spread across four Compositae tribes, and even a separate family, Malvaceae (Funk & Bernays, 2001). In common-garden comparative experimental assays, eastern populations locate, make phloem contact, and settle into long-term feeding on *A. trifida* more efficiently than southwestern populations (Bernays & Funk, 2000; Funk & Bernays, 2001), demonstrating a genetic basis to divergent host-use patterns. Available evidence suggests that these divergent host-use traits have evolved despite ongoing gene flow. Aphids generally have high dispersal rates, there are no disjunctions in the distributions of *Ua* populations, and no differentiation is evident in the handful of mitochondrial and endosymbiont loci analyzed thus far (Funk & Bernays, 2001).

Ua is part of a recent and possibly ongoing adaptive radiation that began 1-2 million years

ago following a recent colonization of North America (Gill, Chapter 1; Moran *et al*, 1999). Species in the clade are highly specific to various Compositae genera, primarily in the tribes Heliantheae and Astereae, though closely related species rarely utilize the same hosts (Moran, 1986). Species distributions are broad and largely coincident in North America, generally matching that of their hosts, though Pleistocene glaciation may have temporarily interrupted gene flow (Gill, Chapter 1). These facts are consistent with the hypothesis that host-associated ecological divergence may have contributed to diversification in the *Ua* clade, rather than emerging subsequent to allopatric speciation events, for example.

Genomic patterns of ecological divergence

To characterize patterns of differentiation across the genome, we distinguish four broad scenarios reflecting different potential evolutionary histories. First, variation at loci across the genome may be undifferentiated in all populations, reflecting an unstructured population with gene flow pervasive enough to overwhelm any locally-imposed selection. Second, variation among populations may be structured by the neutral process of isolation by distance (ID), reflecting a balance between mutation rate, effective population size, and migration (Wright, 1943). Under this scenario, differentiation of homologous loci is expected to increase as geographic distance increases, regardless of differences in local host environment.

Third, a subset of loci may be isolated by ecology (IE)—highly differentiated among populations experiencing divergent ecological selection, but highly similar among same-ecology populations; the extent and size of these "islands of divergence" depends on the number of loci under selection as well as the strength of selection (Nosil *et al*, 2009; Sexton *et al*, 2013; Shafer & Wolf, 2013). IE loci in this study correspond to those that are consistently differentiated according to specialist versus generalist host-breadth. Finally, the latter two scenarios may operate simultaneously, such that a majority of loci reflect ID, while a subset of functional loci reflect IE. In this case, the genome-wide effects of ID must be accounted for in order to identify IE loci that are under divergent selection.

Candidate loci

We take both *a posteriori* (transcriptome scan) and *a priori* (targeted) approaches to characterize the genomic architecture of divergent host-breadth among *Ua* populations. We

define genomic architecture narrowly, in accordance with the scope of this study: the number, diversity, and functional identity (i.e. Gene Ontology categorization) of candidate loci associated with variation in host-breadth. To identify candidate functional loci, we examine *Ua* expressed transcripts with high levels of differentiation among populations (i.e. F_{ST} outliers), and with evidence of positive selection, based on the rate of non-synonymous substitutions. Candidate loci are those that are outliers according to multiple criteria and are consistently differentiated in different-breadth comparisons and consistently similar in same-breadth comparisons (i.e. IE).

A growing body of literature describes a handful of large gene families that have diversified under selection from divergent host plant cues. We focus on these *a priori* gene families—chemosensory proteins and receptors, effector genes, cytochrome p450, detoxification enzymes, and genes expressed in the salivary glands—as distinct units in all analyses. These gene families are of particular interest for ecological divergence because aphids, like many herbivorous insects, reproduce on their host plants. A degree of assortative mating may thus be inherent in divergent host-use traits. Divergence of host-associated loci among IE populations could therefore play a mechanistic role in promoting reproductive barriers, although this notion is largely untested in intra-specific systems. We test the hypothesis that *a priori* host-associated loci show elevated sequence divergence (F_{ST}) and rates of non-synonymous substitution (K_a/K_s) in different-breadth comparisons, relative to genome-wide averages.

We identify these *a priori* candidate loci based on *Uroleucon* homologs of *A. pisum* gene families that are known to respond to host plant cues (Gill, Chapter 2). These gene families include chemosensory loci—including olfactory, gustatory and ionotropic receptor proteins as well as olfactory binding proteins—which mediate host plant selection (Bernays & Chapman, 1994) and are known to evolve rapidly at the sequence and expression levels under selection from novel host ecologies (Kopp *et al.,* 2008; McBride, 2007). Effector genes, including ion channel proteins, are broadly defined as any insect protein (or other functional molecule) that alters host cell structure or function (Hogenhout & Bos, 2011). Many effectors are secreted by salivary glands into the host plant phloem during feeding, and we focus as well on proteins that constitute salivary secretions (Carolan *et al*., 2011; Carolan *et al*, 2009). We also include cytochrome p450 genes, well-described effectors involved in detoxification of plant secondary defensive compounds that have shown greater divergence than other genes in comparisons of closely related species with divergent host use (Berenbaum, 2002).

METHODS

Aphid sampling and sequencing strategy

A complete description of population sampling, RNA extraction, library preparation and sequencing is provided elsewhere (Gill, Chapter 2). Briefly, multiple aphid colonies were sampled from each of five populations representing a large portion of the North American range of *U. ambrosiae.* In this study, we consider four of these populations, constituting two pairs of natural replicates: two eastern U.S. populations that feed on *I. frutescens*, and two generalist populations, one from Arkansas, the other from Arizona and New Mexico (Table 1, Figure 1). Colonies from a given population were pooled in order to obtain estimates of allele frequency, as in several other studies (Harris *et al.,* 2013; Konczal *et al.* 2014). As specialists, all colonies from the two eastern populations were collected feeding on *I. frutescens*. Pooled colonies from the southwestern populations represent collections from multiple host plants (given in Table 1). Host range was denoted as specialist or generalist based on observation of colonization patterns at all sampling localities as well as published host records (Funk & Bernays, 2001; Moran, 1984; Gill, personal observation).

The *Ua* transcriptome assembly described by Gill (Chapter 2) was used as the reference for all SNP analyses in this study. Calculation of K_a/K_s values requires alignment of coding sequences derived independently from each population. We therefore used Trinity with default settings to assemble the raw reads from each population into four population-specific transcriptome assemblies (Haas *et al.,* 2013).

Clustering assembly into isogenes

To generate a set of reference sequences for comparative analysis of polymorphism within and between populations, we filtered *Ua* contigs (Gill, Chapter 2) into a non-redundant set of sequences. In order to apply this filter, USEARCH was used to cluster all contigs sharing an identity threshold of 0.90 (Edgar, 2010). The longest contig associated with each cluster—hereafter referred to as an isogene—was retained. This resulted in a reference transcriptome composed of 81,552 *Ua* isogenes. We chose to impose a strict filter against redundancy in order to create a conservative reference set that would maximize our ability to identify high-confidence SNPs. With a stringent filter, we exclude contigs that may have been misassembled due to false

variation introduced by sequencing errors. Trinity, as a "splice-aware" *de novo* assembler, processes reads to reconstruct multiple possible transcript sequences, including putative paralogs and alternate splice forms (Haas *et al*., 2013). This results in an assembly with very high or complete sub-sequence identity among various subsets of contigs, due to close paralogy or alternately spliced exons. While the presence of both splice forms and paralogs in the assembly reflects biological reality and is helpful to analyze gene expression and alternative splicing, such redundancy is undesirable in a set of reference sequences intended for polymorphism detection, since it creates multiple alignment targets for identical reads (Konczal *et al.*, 2013). Thus, though filtering may have the effect of removing some functionally relevant paralogs and alternate splice forms from the study, we accept that cost in order to minimize the likelihood of accepting false-positive SNPs.

Each isogene represents a "true" assembled sequence—as opposed to a consensus sequence—allowing us to transfer annotations based on the complete, unfiltered *Ua* assembly (Gill, Chapter 2). Similarly, we retain the coding sequences (CDS), *i.e.* open reading frames, previously inferred (Gill, Chapter 2). All GO enrichment analyses conducted in the study were tested using Fisher's exact test with a $p < 0.05$ cutoff, against the null hypothesis that the test group has no different proportional representation of GO terms than the reference group (i.e. the full set of annotated isogenes), as implemented in Blast2GO (Conesa *et al.,* 2005).

SNP calling, allele frequency estimation, calculation of F_{ST}

To identify SNPs within and between populations, each read pool was aligned independently with default settings to the reference isogene set using bowtie2, which is optimized for short-read mapping (Langmead & Salzberg, 2012). We used the samtools software package to remove ambiguously mapped reads (*view* command, with mapping quality threshold = 20) and to compile alignments from all four populations into the mpileup format (Li *et al*., 2009). Samtools integrates quality scores and uses Bayesian inference to make SNP calls with high confidence (Nielsen *et al*, 2011).

For additional quality-control for SNP calls and to calculate allele frequency differences, we passed the compiled alignments to PoPoolation2, software designed for probabilistic SNP calling based on modeling sequencing error and uneven sampling resulting from sequencing pooled DNA (Kofler *et al*, 2011; Futschik & Schlotterer, 2010). Pooling multiple genotypes prior to

sequencing is a commonly used method of efficiently estimating allele frequencies in non-model systems, for which the costs of individually preparing sequencing libraries at sufficient breadth and depth to calculate individual-based genotypes remains prohibitive (Cutler & Jensen, 2010). To minimize the likelihood of accepting sequencing errors as valid SNPs and to avoid biased estimates due to very low coverage, we considered only positions with a minimum coverage of ten reads for all populations, and a minimum allele count of two in at least one population. Hence our final SNP set is a conservative estimate of the true SNPs segregating within and between *U. ambrosiae* populations. Pairwise allele frequency differences and F_{ST} values were calculated among all four populations for each SNP, and also for non-overlapping sliding windows 300 nucleotides long. Fisher's exact test was used to detect loci with significantly different allele frequencies in different pairwise comparisons, as implemented in PoPoolation2.

The F_{ST} calculations implemented in PoPoolation2 are based on allele frequency estimates. Although allele frequency estimates are common from pooled genomic DNA, it is important to note that our sequence reads derive from expressed transcripts that, unlike genomic DNA, may lead to estimates of allele frequencies that are biased either by allele-specific expression or among-individual differences in expression (Garvin *et al*, 2010). However it has recently been shown empirically that variant frequencies calculated from pooled RNA-seq data are highly correlated with true allele frequencies (Konczal *et al*., 2014), possibly because relatively few loci show significant expression differences among individuals. Konczal *et al.* (2014) explicitly tested the accuracy of allele frequency estimates derived from pooled RNA-seq in the absence of a reference genome by sequencing vole liver transcriptomes both individually and as pools. They found that estimates from pooled RNA are highly correlated to "true" frequencies as measured by individual-level sequencing, and are comparable to those derived from pooled genomic DNA. Additionally, cDNA libraries were normalized using a digestion method that preferentially removes the most abundant sequence templates, reducing any effects that may result from drastically different sequencing sample sizes across loci (Gill, Chapter 2). Nevertheless, we enforce a high frequency difference threshold to regard a particular SNP as differentiated in any pairwise comparison. We also take our population genetics analyses as preliminary until candidate isogenes from this study can be validated using targeted re-sequencing based on genomic DNA.

K_a/K_s scan

To identify population-specific coding sequences, we individually assembled the reads from each population into population-specific contigs using Trinity (Haas *et al*., 2013). Homologous sequences were identified in each population based on reciprocal best matches in pairwise comparisons using USEARCH (Edgar, 2010). Homologous sequences were then used to generate pairwise alignments of all isogenes for each population comparison. These alignments, along with protein sequences based on inferred open reading frames, were passed to parAT, which generates parallel alignments of protein sequences and nucleotide coding sequences for each isogene with an ORF (Zhang *et al.,* 2012). Protein and nucleotide alignments were used to calculate K_a/K_s values for each pairwise alignment with KaKs Calculator (Zhang *et al*., 2006) using the GY method, which implements a maximum likelihood approach accounting for transition/transversion rates and nucleotide frequencies. Significance of K_a/K_s values was determined with Fisher's exact test under the alternative hypothesis that the non-synonymous substitution rate is different than the synonymous substitution rate, as implemented in KaKs Calculator.

To test significant differences in the frequency distributions of K_a/K_s and F_{ST} among populations, we use the Kolmogorov-Smirnov test, which does not assume normally distributed data. To test for significant differences of mean values among populations we use the Mann-Whitney U test, a version of the t-test that performs well on non-normally distributed data.

RESULTS

Population-specific transcriptome assembly

Results for population-specific Trinity assemblies are given in Table 2. Assemblies are broadly similar in numbers of assembled contigs and total assembly size. Open reading frames were inferred for each assembly, resulting in around 30,000 coding sequences per population.

Host-associated genes

In total 419 *Ua* sequences were assigned orthology to host-use associated pea aphid gene families (Gill, Chapter 2), which are generally computationally annotated based on inferred homology to other annotated insect genes (IAGC, 2010; Legeai *et al.,* 2010). These include chemosensory proteins and receptors (29 *Ua* sequences matched a database of 105 pea aphid genes, a match rate of 27.6%), p450 detoxification enzymes (50/85; 58.9%), salivary secretion constituents (163/279; 58.4%), circadian clock genes (37/13; 280%), and ion channel genes (140/93; 151%). We compare this set of genes to our outlier results to detect signs of divergent host-associated evolution at loci of *a priori* interest.

SNP identification

Mapping of reads from each population to the set of reference *Ua* isogenes resulted in a mean sequencing depth of 44.9 reads per polymorphic site across 9340 unigenes (Figure 2). Nearly 75,000 sites had coverage of less than ten reads, indicating that read coverage did not approach saturation of expressed sequences and that greater sequencing effort would uncover a greater degree of polymorphism than is considered in the present study. At low sequencing coverage it is difficult to distinguish true SNPs from the effects of sequencing, assembly and alignment error. Sites with minimum coverage (C_{min}) of less than ten reads in at least one population were discarded, leaving 296,651 sites spanning 8630 unigenes with mean sequencing depth of 54.5 reads. Even when read coverage is sufficient, allele frequency estimation is unreliable if based on only a single count of a minor allele due to the possibility of sequencing error (Van Tassell *et al.,* 2008). A minimum minor allele frequency was therefore set to $P > 0.1$. Setting $C_{min} \geq 10$ and $P > 0.1$ effectively enforces that a minor allele must be supported by a minimum read count of 2, excluding singletons that may otherwise be counted as false positives.

Using a strict quality filter risks rejecting true positives, but these would be at the lowest allele frequencies and therefore of minor interest for the goal of identifying functionally relevant loci.

Imposing this filter resulted in 39,245 SNPs on 6,240 isogenes, out of over 4 million total nucleotide sites (including monomorphic sites) in the non-redundant *Ua* reference isogene set. This represents a rate of polymorphism of 0.96%, comparable to the 0.75% found in *Drosophila simulans* (Begun *et al*., 2007). On average each polymorphic transcript has 6.3 SNPs, although a fifth of all transcripts have only a single SNP (Figure 3). These results are comparable to those found in other *de novo* insect transcriptome studies. For example, 38,141 SNPs were found in 2,907 different transcripts for the ground beetle *Pogonus chalceus* (Van Belleghem *et al*, 2012).

SNP variation across populations

To characterize genetic variation among *Ua* populations, we calculated the pairwise differentiation index, $D = | P_{pop1} - P_{pop2} |$, for all transcripts based on the per-locus allele frequency *P* (Andrés *et al*, 2013). Significance of D values is based on Fisher's exact test against the null hypothesis that $P_{pop1} = P_{pop2}$ using a cutoff of $p<0.05$, corresponding to a false discovery rate of $q<0.01$. Only SNP sites meeting this criterion were considered. The distribution of *D* values is distinctly bimodal (Figure 4). Over 80% of SNPs have low allelic divergence in all pairwise comparisons ($D < 0.2$). A small peak above D=0.1 is an artifact of our filtering criterion of a minimum minor allele frequency of 0.1 (which is equivalent to a minimum minor allele count of 2). This eliminates *P* values within the interval of ($0 < D < 0.1$), resulting in what appears to be a spike in allele frequency differences.

A number of isogenes are highly diverged at $D > 0.9$, suggesting these populations are at or approaching fixation for alternate alleles. This includes 641 SNPs distributed over 514 unique isogenes that are consistently differentiated in specialist-generalist comparisons and undifferentiated ($D < 0.2$) in same-breadth comparisons, representing 8.24% of the total number of polymorphic isogenes. A subset of these, covering 2.31% of all polymorphic loci—equivalent to 0.4% of all SNPs—are fixed for divergent alleles ($D = 1$) in all generalist versus specialist comparisons. The 2.31% of SNPs with nearly fixed differences approaches the 4% found in a comparison of two closely related cricket species (Andrés *et al*., 2013), and is comparable to the 7.5% found in the mosquito *Anopheles gambiae,* among M and S forms, which can be considered incipient species (Lawniczak *et al.,* 2010). On average, *D* is significantly lower in

same-breadth pairwise comparisons than generalist-specialist comparisons (0.10 vs. 0.17, $p << .001$).

We compared the average *D* value for host-associated genes to a test set of an equivalent number of sequences selected at random and resampled 100 times from the reference transcriptome (Table 3). Ion channel genes had an average D in generalist-specialist comparisons significantly above random isogenes ($D_{avg} = 0.29$, $p = .0003$). D values for these two groups in same-breadth comparisons were not significantly different. COO2 genes also had significantly elevated allele frequency difference in generalist-specialist comparisons ($D_{avg} = 0.545$, $p = .035$). Salivary gland genes had the next highest average D value at D = 0.165. This is equivalent to the value for random genes, although the standard deviation (sd = 0.258) is high, suggesting that a subset of these genes is highly differentiated. Of the 514 transcripts with D > 0.9 in all specialist-generalist comparisons, ten are members of *a priori* host associated gene families, including 7 salivary genes, 2 ion channel genes, and one p450 gene.

F_{ST} scan

F_{ST} was calculated across each isogene in non-overlapping windows of 300 nucleotides based on $F_{ST} = \sigma_S^2/\sigma_T^2$, where σ_S^2 is the variance in the frequency of alleles in subpopulations, and σ_T^2 is the variance of allele frequencies in the total population. F_{ST} is calculated directly from allele frequency estimates, and was only calculated for SNPs with significant allele frequency differences as described above. Only windows $C_{min} \geq 10$ across the entire window were accepted, resulting in 6242 total windows in 2,777 isogenes. Our main finding is that a larger number of isogenes have elevated F_{ST} (>0.6) in pairwise comparisons of specialist and generalist populations compared to same-breadth populations. The average pairwise F_{ST} value across all loci in specialist-generalist comparisons is 0.19, significantly higher than the 0.14 for same-breadth comparisons ($p < 2.2 \times 10^{-16}$). The distribution of pairwise F_{ST} window values (Figure 5) shows that the elevated specialist-generalist average F_{ST} value is due to a higher density of isogenes from $F_{ST} = 0.5 – 1.0$ in specialist-generalist comparisons (warm colors in Figure 5). This difference is highly significant, as equivalence of specialist-generalist compared to same-breadth distributions was rejected with confidence ($p= 6.024 \times 10^{-11}$).

Despite divergent distributions at elevated F_{ST} levels, the vast majority of polymorphic sites have low F_{ST} values—they are not more differentiated between populations than they are

within populations. This mass of undifferentiated loci is represented by the peaks toward the left side of Figure 5. F_{ST} values in this range show no particular pattern in same-breadth versus different-breadth comparisons. Despite highly different mean values, similar F_{ST} levels specifically at the low ($F_{ST} < 0.2$) end of the distribution are supported by nearly identical median values when comparing the pooled set of all specialist-generalist F_{ST} values to the respective same-breadth comparison (0.1186 vs. 0.1184). Our interest is in isogenes that are highly differentiated (*i.e.* outliers) relative to the transcriptome-wide distribution, under the presumption that at least some portion of loci with allele frequencies that are widely divergent in multiple different-host comparisons may be functionally responding to divergent selection or host-use ecology more generally.

The number of transcripts in same-breadth comparisons (*i.e.* NE vs. AC; AZ vs. AR) declines to near zero above an F_{ST} of 0.4—95% of same-breadth F_{ST} values are at $F_{ST} < 0.32$. 99% of these values are distributed at $F_{ST} < 0.47$. In contrast, the 95th and 99th percentiles in specialist-generalist comparisons (*i.e.* NE vs AR, NE vs. AZ, AC vs. AR, AC vs. AZ) are F_{ST} = 0.63 and 0.86, respectively. A fifth (563/2777) of all F_{ST} windows in generalist-specialist comparisons have $F_{ST} > 0.6$. Of these, we focus on 103 (3.7%) isogenes that are differentiated in all four generalist-specialist comparisons, but are simultaneously undifferentiated in both same-ecology comparisons. Only one sequence has a pairwise F_{ST} value greater than 0.6 in both same-range comparisons.

According to our BLAST query of all *Ua* isogenes to the NCBI non-redundant protein database, the majority (83%) of these F_{ST} outliers are most similar to pea aphid genes, 11% to *Buchnera* genes, and the remaining have their closest match in other insect species. When reduced to the most specific terms, 49 gene ontology (GO) terms were significantly over-represented among the 103 outlier isogenes compared to the GO term distribution across the entire isogene reference set (Fisher's exact test, $p < 0.05$). This suggests some redundancy in potential biological functions of outliers, though only a few GO categories are represented by multiple F_{ST} outliers. These include two isogenes involved in beta-alanine metabolism, which produces metabolites present in honeydew, and two involved in binding magnesium ions, which are among the most versatile biochemical cofactors (Dudev & Lim, 2003). However, the majority of outlier isogenes are annotated with distinct GO terms, suggesting that isogenes associated with a relatively wide variety of functional traits are consistently differentiated in

different-breadth populations. We also compared F_{ST} outliers to the 419 *a priori Ua* host-use associated transcripts discussed above—two are F_{ST} outliers, both salivary gland genes.

K_a/K_s scan

We tested all polymorphic coding sequences across all four populations for signs of positive selection based on the pairwise ratio of non-synonymous substitutions to synonymous substitutions (K_a/K_s). The K_a/K_s ratio was first developed to apply to distantly related taxa, though it has increasingly been applied to closely related species (Andrés *et al*., 2013; Brieuc & Naish, 2011) as well as conspecific populations (Barreto *et al*, 2011; Harris *et al*, 2013). It is particularly useful in the context of non-model transcriptomics because it is expected to perform well for SNPs derived from pooled samples (Baldo *et al*, 2011).

K_a/K_s values are based on estimating non-synonymous (K_a) and synonymous (K_s) substitution rates for alignments of homologous coding sequence for each pairwise population comparison. To determine which coding sequences had significantly different substitution rates, we calculated the distribution of q-values based on p-values from Fisher's exact test of the null hypothesis that $K_a = K_s$. The q-value distribution is based on the false discovery rate associated with p-value cutoffs, providing an estimate of the expected number of false positives among a set of features deemed significant at a given cutoff. Analysis of the distribution of q-values indicates that a large proportion of the examined loci have significantly different K_a and K_s values (Figure 6). At a significance cutoff of $q < 0.006$ (corresponding to $p < 0.1$), which is used here, only 0.6% of significant features is expected to be a false positive. Based on this cutoff, 2043 of 4737 coding sequences are accepted as having significant K_a/K_s ratios, and only around 12 are expected to be false positives. There is only a slight correlation between K_a/K_s and length ($R^2 = 0.002$, $p<0.01$), suggesting that the calculating method is not biased by sequence length despite the greater number of substitutions likely in longer sequences (Figure 7). This is as expected, since KaKs Calculator corrects for sequence length variation using a maximum likelihood model that takes sequence length into account in estimating substitution rates (Zhang *et al,* 2006).

K_a/K_s tests are commonly used to infer positive selection, though they should be interpreted cautiously (Ellegren, 2008). A strict cutoff for evidence of selection is sometimes taken to be $K_a/K_s > 1$, indicating isogenes for which the majority of substitutions are non-synonymous. However, a given isogene may have a number of SNPs, and selection may

potentially act on a single position, hypothetically resulting in K_a/K_s values less than one, but still reflecting positive selection. In closely related populations, few mutations may have had time to rise to appreciable frequency regardless of codon site—the existence of one or two non-synonymous substitutions may be of functional relevance even if there are several more silent substitutions that would drive down the K_a/K_s value. Thus lower K_a/K_s cutoffs have also been utilized as a way to reliably detect selection (*e.g.* Swanson, 2004).

K_a/K_s values for each coding sequences were calculated for all six pairwise comparisons among the four populations. The large majority of isogenes did not exhibit evidence of elevated non-synonymous substitution rates, based on the mean K_a/K_s values of all pairwise comparisons, regardless of host ecology (mean $K_a/K_s = .053$). This is somewhat lower than found in other attempts at intra-specific K_a/K_s scans (0.12 among populations of the copepod *Tigriopus californicus* (Barreto *et al*., 2011); and 0.28 among urban and rural populations of the white-footed mouse *Peromyscus leucopus* (Harris *et al*., 2013). The concentration of low K_a/K_s values appears as a peak below $K_a/K_s < 0.3$ in Figure 8, which shows the density distribution of isogenes across K_a/K_s values.

Comparing values for all isogenes across all pairwise comparisons, we identify 11 loci with K_a/K_s values > 0.4 in all each of the four specialist-generalist comparisons, compared to 4 transcripts with $K_a/K_s > 0.4$ in both same-breadth comparisons. $K_a/K_s > 0.4$ is a relaxed cutoff compared to many other studies and as a result, we do not take it as firm evidence that identified loci are diverging according to natural selection. Rather, our interest is in identifying loci that are most likely to be functionally diverging among different-host-breadth populations, and we use this cutoff as a way of casting attention on the subset of loci with the most elevated rates of non-synonymous substitution.

DISCUSSION

We characterized genome-wide patterns of differentiation across expressed sequences within and among populations of *Ua*, with the purpose of characterizing patterns of sequence divergence throughout the transcriptome. Our main results are, first, that a subset of *Ua* isogenes is isolated by ecology, *i.e.* consistently differentiated when comparing generalist versus specialist populations, but not when those same loci are compared in specialist versus specialist or generalist versus generalist populations. Although this does not guarantee that host-breadth is the ecological cause of differentiation (addressed below), it suggests high priority loci for further analysis. Second, we identified four primary candidate isogenes that are consistently isolated by ecology, with elevated D, elevated F_{ST}, and elevated K_a/K_s values in all generalist-specialist comparisons. We also found evidence for a number of secondary candidate genes that pass two out of these three screens, including several members of two *a priori* host-associated gene families: ion channel genes and salivary gland genes.

Genomic differentiation

Identifying candidate genes that are responding to divergent ecological selection requires distinguishing patterns of adaptive evolution from those expected under neutral processes. To address this, our initial goal was to analyze patterns of genetic variation within and among *Ua* populations to evaluate support for one of three ecological scenarios. Populations may exhibit little population structure, with abundant gene flow attenuating among-population divergence throughout the genome. They may be isolated by distance, leading to neutral divergence among a subset of loci. Or, possibly in combination with isolation by distance, populations may be isolated by ecology (IE), with a subset of loci showing evidence for non-neutral, adaptive evolution. We find evidence suggesting that a subset of loci among *Ua* populations is isolated by ecology.

Previous population-level analysis of *Ua* revealed strikingly low levels of polymorphism across populations (Funk *et al*, 2000). This is consistent with widespread gene flow among populations across *Ua*'s range, though evidence was limited to analysis of three *Buchnera* loci and one mitochondrial locus. We find clear evidence for population structure among *Ua* populations. The distribution of pairwise allele frequency differences and F_{ST} statistics suggests

that *Ua* populations exchange genes, but not enough to preclude consistent population structure at a subset of loci. Evidence for population structure is provided by the bimodal distribution of D (pairwise per-locus allele frequency differences), which shows a peak at D > 0.9, indicating a high degree of differentiation across several populations (Figure 4).

Population-level variation may result from adaptive divergence, but also from neutral processes—in the latter case, both isolation by distance and Pleistocene isolation followed by secondary contact are plausible for *Ua* (Gill, Chapter 1). Differentiation among *Ua* populations is unlikely to be attributable solely to geographic distance, since mean pairwise allele frequency difference and mean pairwise F_{ST} are not significantly correlated with geographic distance among the population sampled in this study (Figure 9), suggesting that alternate forces are also driving genomic divergence. A larger point is that neutral processes are expected to affect loci throughout the genome uniformly. Yet all pairwise comparisons show similar median values, around $F_{ST} = 0.12$, despite the presence of multiple highly differentiated loci that are specific, respectively, to generalist and specialist populations. Thus, although the vast majority of loci in all comparisons have similar F_{ST} profiles, a number of isogenes are highly differentiated uniquely in different-breadth comparisons.

We use a multiple comparisons approach to identify these isogenes by searching for loci that are differentiated due to adaptive rather than neutral causes among populations (in IE), *i.e.* strongly differentiated in different-breadth but not in same-breadth comparisons. We found strong support for greater levels of genomic differentiation between populations with different host breadth, compared to populations with the same host breadth. Same-breadth pairwise comparisons have a lower average allele frequency difference (0.10) than different-breadth comparisons (0.17; Mann-Whitney test, $p << .001$; Figure 4). As expected, F_{ST}, which is related to D in that both are based on allele frequency estimates, shows a similar pattern, with elevated F_{ST} among a subset of loci in specialist versus generalist comparisons (Figure 5).

The distribution of these statistics supports the view that most polymorphisms in *Ua* populations are segregating at synonymous sites and are undifferentiated in relation to both geography and host use. A small subset of isogenes, however, are outliers that are both highly differentiated and show elevated rates of non-synonymous substitutions in multiple comparisons of different-breadth populations. We use the distributions of D and F_{ST} to assess population structure across transcriptomic loci among populations, but also as a screen for genes with

possible functional roles in host-use divergence. To further this goal, we also consider K_a/K_s values, which are not based on allele frequency differences and therefore provide an independent screen to detect adaptively diverging loci. The inferred functional roles of isogenes that are differentiated specifically in different-breadth comparisons provide further evidence that differentiation of these loci may be a result of adaptive evolution in accordance with divergent host ecology.

Transcriptome scans for host-associated candidate genes

A main goal of this study is to identify candidate genes that may be playing a functional role in adaptive diversification among populations with divergent host-use traits. To qualify as candidates, isogenes must be consistently differentiated in ecologically divergent populations, as discussed above, and also show evidence of positive selection (*i.e.* elevated K_a/K_s).

We identified a relatively large number of loci identified as D and F_{ST} outliers, and a small number of K_a/K_s outliers (514, 103 and 11, respectively). To identify candidate loci, we targeted the intersection of the transcript sets identified using each scanning method (Figure 10). First, we examine the intersection of isogenes with $K_a/K_s > 0.4$ and $D > 0.9$ in each of the four specialist-generalist comparisons. In intraspecific comparisons it important to consider allele frequency differences, since K_a/K_s calculations assume that alleles are fixed between populations (Kryazhimskiy & Plotkin, 2008). In applying a $D > 0.9$ filter on the K_a/K_s results, we minimize the chance that we erroneously accept elevated K_a/K_s values that are calculated over segregating sites rather than fixed differences.

Nine loci have both $K_a/K_s > 0.4$ and $D > 0.9$ in all specialist-generalist comparisons. When compared with the F_{ST} scan, four transcripts emerge as consistent outliers in all analyses. These loci correspond to an isolation by ecology scenario and meet our criteria for candidate genes (Table 4).

Annotations for each of the four candidate genes suggest plausible functional roles in divergent host-use adaptation. The first, comp40311_c0_seq1, has high similarity to a single pea aphid gene (evalue = 0, identity = 93%), an α-glucosidase gut sucrase gene that is thought to be the dominant and perhaps sole enzyme responsible for sucrase activity in the gut (Price *et al.,* 2007). Gut sucrase enzyme activity is vital to two physiological mechanisms in aphids. In the first, sucrase hydrolyzes sucrose, the dominant source of organic carbon in phloem sap, into

monosochaccharides that are nutritionally available. Sucrose also accounts for the high osmotic pressure of phloem sap, which presents a challenge to aphids that sucrase helps surmount (Fisher *et al*, 1984). Sucrase plays a dominant role in osmoregulation by allowing the assimilation of sucrose across the gut wall, thereby reducing the osmotic pressure inside the gut, facilitating feeding (Price *et al.,* 2007). Our finding that an α-glucosidase-like gene is consistently differentiated in generalists versus specialists, and shows evidence of positive selection, suggests that generalist *Ua* populations may be adapting to differing phloem sucrose profiles expressed by their hosts.

Another candidate, comp44978_c0_seq1, is a high-identify match to succinyl-CoA synthetase (the only hit to the corresponding *Ap* gene has e-value = 0, identity = 93%), the only mitochondrial enzyme capable of producing ATP via substrate level phosphorylation without oxygen, in addition to a role in the citric acid cycle (Sabri *et al.,* 2013). It is unclear to which, if any, host-associated phenotype this gene may contribute, though its presence in pea aphid honeydew suggests that it plays a role in feeding (Sabri *et al.,* 2013). This is further supported by findings that succinyl-CoA is one of just ten proteins down-regulated in the proteome of the green peach aphid *Myzus persicae,* a generalist, when feeding on potato but not other host plants (Francis *et al.,* 2006).

A third candidate, comp47132_c0_seq1, is a nuclear ribonucleoprotein, based on inferred similarity to a pea aphid gene (e-value = 0, identity=91%). Intriguingly, the fourth candidate, comp54019_c0_seq1, is a ribosomal *Buchnera* gene that encodes an enzyme modifying a uridine nucleoside that is a constituent of several tRNA species (Silva *et al.*, 2006). Endosymbionts have been implicated previously in divergent aphid host-use (Leonardo & Muiru, 2003; Mclean *et al.*, 2011; Tsuchida *et al.*, 2011), and orthologs of this particular gene are known among symbiont species (Jiang *et al.,* 2013). These facts are consistent with a potential role for intra-specific divergence of *Buchnera* genes in host-use of *U. ambrosiae,* in this case possibly via translational regulation.

Loci that are outliers in some but not all tests may also be of interest, though further investigation would be required to support a functional role. We annotate one F_{ST} outlier, for example, as a cathepsin B gene (cathepsin B-348). Genes in the cathepsin family are cysteine proteases involved in intestinal digestion of proteins. Though aphids have previously been presumed to express few digestive proteins (instead relying on *Buchnera* endosymbionts),

cathepsin B gene families were found to be amplified in the pea aphid relative to *D. melanogaster*. A number of these cathepsin B genes had high rates of non-synonymous substitutions (Rispe *et al.,* 2007). This pattern may be indicative of adaptation to the phloem environment faced by aphids as they diverged from their common ancestor with other insects. Evidence of divergent selection on cathepsin-B in Ua supports the notion that protease genes are evolving under divergent host-associated selection.

One K_a/K_s outlier is involved in serine-type endopeptidase inhibitor activity, which can allow aphids to detoxify proteinase inhibitors that are common in phloem. These genes have been shown to contribute to host-breadth by enabling herbivorous insects to detoxify PIs from a variety of species (Falco & Silva-Filho, 2003). Serine-type endopeptidase inhibitor genes are part of a lineage-specific expansion in *Uroleucon* relative to *Ap* (Gill, Chapter 2), and evidence that some genes in the family are undergoing divergent non-synonymous evolution in generalist versus specialist populations is suggestive of a functional role in host-use evolution.

a priori *host-associated gene families*

Our results point to an adaptive role for some salivary gland genes, including especially ion channel genes. Seven salivary and two ion channel isogenes (and one cytochrome p450 isogene, which like salivary genes are often mobilized in the phloem) have $D > 0.9$. Two additional genes, one salivary and one p450, have $K_a/K_s > 0.4$, suggesting divergent selection in generalist or specialist populations, or both. Our F_{ST} analysis also revealed two *a priori* genes at $F_{ST} > 0.6$, both salivary genes. One of these salivary genes, comp54057_c0_seq2, is also identified by elevated D.

These results cast new light on findings in *A. pisum*, which shows high levels of differentiation and signs of divergent positive selection among host races at gustatory and odorant receptor loci (Smadja *et al.,* 2012). Since pea aphid host races are intra-specific, divergent selection among chemosensory loci could reasonably be assumed to be an important and perhaps, for aphids at least, ubiquitous mechanism underlying divergent host use traits.

Our results do not support an early role for sequence evolution of chemosensory loci in the evolution of divergent host-use traits. Although we likely do not recover the full set of chemosensory proteins encoded in *Ua,* not one of the 29 *Ua* chemosensory loci we identified was polymorphic. In comparison to the evident purifying selection on the chemosensory genes in

our analysis, the consistent differentiation of a subset of salivary gland and ion channel genes stands out. Smadja *et al.* (2012) specifically targeted chemosensory genes, comparing them to random loci, but did not address the other host-associated gene families discussed here. *Ua* populations are qualitatively less differentiated than *Ap* host races, which are highly genetically structured by host plant, even in sympatry (Frantz *et al.*, 2006). These results point to the possibility that evolution at loci that directly interact with plant phloem may play earlier roles in host-use adaptation than chemosensory loci. This hypothesis predicts that in the pea aphid, salivary gland, ion channel and other genes encoding phloem-mobilized proteins would show greater differentiation and coalesce earlier in time than chemosensory loci.

Evidence that several salivary gland genes are functionally diverging among specialist and generalist populations suggests that divergent phloem profiles of alternate host plant taxa may be driving host-associated differentiation at these loci. One obstacle in understanding the dynamics of incipient or ongoing host-use divergence is that research on divergent ecological adaptation among sympatric or parapatric populations has largely focused on comparing completely or partially reproductively isolated taxa. One objective of the present study is to establish *U. ambrosiae* as a potential model for the study of ecological divergence among intraspecific populations. We demonstrate an approach to identifying candidate loci for intraspecific ecological divergence in a non-model system with no previously existing genomic resources. Yet the candidate genes that we identify based on pooled transcriptome sequencing are derived from only four populations. Direct, population-level resequencing approaches that target the broad set of host-use gene families, along with a reference set of randomly selected and/or neutral loci, are a promising area for future research to confirm the present findings and further characterize the genomic signature of ecological divergence.

Tables

Table 1. Summary of aphid and sequence sampling for each population used in the study.

Sample ID	Sampling localities	No. colonies sampled	Host breadth	Host taxa
NE (northeast)	NY, MA, NH	10	Specialist	*I. frutescens*
AC (mid-Atlantic Coast)	VA, NJ	5	Specialist	*I. frutescens*
AR (Arkansas)	AR	6	Generalist	*Ambrosia, Bidens, Eupatorium*
SW (southwest)	AZ, NM	15	Generalist	*Ambrosia, Viguiera, Tithonia, Heterotheca*

Table 2. Results for population-specific Trinity transcriptome assemblies. The reads used to generate these assemblies are the same as those used for the *Ua* transcriptome, although in this case only reads derived from each population were used for the respective assemblies. n = number of sequences; N50 = over half the assembled bases are in contigs of this length or longer; max = maximum contig length; sum = total length of assembly in megabases; sum = total length of all assembled sequences.

	n	N50	sum
Nucleotide assembly			
NE	43,140	983	2.85E+07
AC	46,277	962	3.08E+07
AR	42,153	951	2.74E+07
SW	34,474	1036	2.37E+07
Open reading frames			
NE	30,095	933	1.91E+07
AC	32,069	927	2.04E+07
AR	29,185	945	1.85E+07
SW	27,428	1249	2.18E+07

Table 3. T-tests for allele frequency differences between *a priori* gene families and a randomly selected set of isogenes (sampled with replacement 100 times). Ion channel genes and COO2 homologs both show significantly elevated D in generalist-specialist comparisons. D_{avg} refers to the average of all pairwise frequency differences. SD is the standard deviation. A single D_{avg} and SD are calculated for pairwise comparisons in all four generalist-specialist comparisons (top) and both same-breadth comparisons (bottom), respectively. P is the p-value for tests comparing D_{avg} for the given gene family to D_{avg} for the random set.

	D_{avg}	SD	P
Generalist - specialist			
COO2	0.545*	0.062	0.035
p450	0.155	0.245	0.350
Circadian	0.132	0.195	0.095
Ion channel	0.286**	0.357	0.0003
Salivary	0.165	0.258	0.505
Chemosensory	0	0	N/A
Random	0.169	0.251	
Same-breadth			
COO2	0.055	0.077	0.529
p450	0.093	0.124	0.284
Circadian	0.077	0.119	0.170
Ion channel	0.107	0.148	0.844
Salivary	0.097	0.141	0.080
Chemosensory	0	0	N/A
Random	0.104	0.148	

Table 4. Candidate genes. Each of these isogenes has $K_a/K_s > 0.4$, $F_{ST} > 0.6$, $D > 0.9$ in all four generalist-specialist comparisons.

Sequence name	Sequence desc.	GO terms
comp40311_c0_seq1	sucrase precursor	GO:0005975,GO:0043169,GO:0003824
comp44978_c0_seq1	succinyl-CoA synthetase	GO:0016874, GO:0005524, GO:0008152
comp47132_c0_seq1	heterogeneous nuclear ribonucleoprotein	GO:0003676, GO:0030529, GO:0000166
comp54019_c0_seq1	50s ribosomal protein l2 (Buchnera)	GO:0003729, GO:0003735, GO:0015935, GO:0015934, GO:0016740, GO:0006412, GO:0019843, GO:0042254

CHAPTER 4 REFERENCES

Andrés, J. A. J., Larson, E. L. E., Bogdanowicz, S. M. S., & Harrison, R. G. R. (2013). Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics, 193(2), 501–513.

Baldo, L., Santos, M., & Salzburger, W. (2011). Comparative transcriptomics of eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. Genome Biology and Evolution, 3, 443-455.

Barreto, F. S. F., Moy, G. W. G., & Burton, R. S. R. (2011). Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus*. Molecular Ecology, 20(3), 560–572.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biology, 5(11), 2534-2559.

Berenbaum, M. (2002). Postgenomic chemical ecology: from genetic code to ecological interactions. Journal of Chemical Ecology, 28(5), 873–896.

Bernays, E. A., & Chapman, R. F. (1994). Host-plant selection by phytophagous insects. Springer.

Bernays, E. A., & Funk, D. J. (2000). Electrical penetration graph analysis reveals population differentiation of host-plant probing behaviors within the aphid species *Uroleucon ambrosiae*. Entomologia Experimentalis et Applicata, 97(2), 183–191.

Brieuc, M. S. O., & Naish, K. A. (2011). Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources. Molecular Ecology Resources, 11, 172–183.

Carolan, J. C., Caragea, D., Reardon, K. T., Mutti, N. S., Dittmer, N., Pappan, K. (2011). Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): A dual transcriptomic/proteomic approach. Journal of Proteome Research, 10(4), 1505–1518.

Carolan, J. C., Fitzroy, C. I. J., Ashton, P. D., Douglas, A. E., & Wilkinson, T. L. (2009). The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. Proteomics, 9(9), 2457–2467.

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? Genetics, 186(1), 41–43.

Davey, J. W., & Blaxter, M. L. (2011). RADSeq: next-generation population genetics. Briefings in Functional Genomics, 9(5), 416–423.

Dudev, T., & Lim, C. (2003). Principles governing Mg, Ca, and Zn binding and selectivity in proteins. Chemical Reviews, 103(3), 773–788.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460–2461.

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and plants: a study in coevolution. Evolution; International Journal of Organic Evolution, 18(4), 586–608.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. Molecular Ecology, 17(21), 4586–4596.

Falco, M. C., & Silva-Filho, M. C. (2003). Expression of soybean proteinase inhibitors in transgenic sugarcane plants: effects on natural defense against *Diatraea saccharalis*. Plant Physiology and Biochemistry, 41(8), 761–766.

Fisher, D. B., Wright, J. P., & Mittler, T. E. (1984). Osmoregulation by the aphid *Myzus persicae*: A physiological role for honeydew oligosaccharides. Journal of Insect Physiology, 30(5), 387–393.

Francis, F., Gerkens, P., Harmel, N., Mazzucchelli, G., De Pauw, E., & Haubruge, E. (2006). Proteomics in *Myzus persicae*: effect of aphid host plant switch. Insect Biochemistry and Molecular Biology, 36(3), 219–227.

Frantz, A., Plantegenest, M., Mieuzet, L., & Simon, J. C. (2006). Ecological specialization correlates with genotypic differentiation in sympatric host-populations of the pea aphid. Journal of Evolutionary Biology, 19(2), 392–401.

Funk, D. J., Nosil, P., & Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. Proceedings of the National Academy of Sciences, 103(9), 3209–3213.

Funk, D., & Bernays, E. (2001). Geographic variation in host specificity reveals host range evolution in *Uroleucon ambrosiae* aphids. Ecology, 82(3), 726-739.

Funk, D., Helbling, L., & Wernegreen, J. (2000). Intraspecific phylogenetic congruence among multiple symbiont genomes. Proceedings of the Royal Society B: Biological Sciences, 267(1461), 2517-2521.

Futschik, A., & Schlotterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics, 186(1), 207–218.

Futuyma, D. J., & Agrawal, A. (2009). Macroevolution and the biological diversity of plants and herbivores. Proceedings of the National Academy of Sciences, 106(43), 18054-18061.

Garvin, M., Saitoh, K., & Gharrett, A. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. Molecular Ecology Resources. 10(6), 915-934.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494–1512.

Harris, S. E., Munshi-South, J., Obergfell, C., & O'Neill, R. (2013). Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York metropolitan area. PLoS ONE, 8(8), e74938–e74938.

Hogenhout, S. A., & Bos, J. I. (2011). Effector proteins that modulate plant–insect interactions. Current Opinion in Plant Biology, 14(4), 422–428.

Janz, N., & Nylin, S. (2008). The oscillation hypothesis of host-plant range and speciation. Specialization, speciation, and radiation: the evolutionary biology of herbivorous insects. University of California Press, Berkeley, 203–215.

Jiang, Z. F., Xia, F., Johnson, K. W., Brown, C. D., Bartom, E., Tuteja, J. H. (2013). Comparison

of the genome sequences of "*Candidatus Portiera aleyrodidarum*" primary endosymbionts of the whitefly *Bemisia tabaci* B and Q biotypes. Applied and Environmental Microbiology, 79(5), 1757–1759.

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics, 27(24), 3435–3436.

Konczal, M., Koteja, P., & Stuglik, M. T. (2014). Accuracy of allele frequency estimation using pooled RNA-Seq. Molecular Ecology, 14: 381-392.

Kopp, A., Barmina, O., Hamilton, A. M., Higgins, L., Mcintyre, L. M., & Jones, C. D. (2008). Evolution of gene expression in the *Drosophila* olfactory system. Molecular Biology and Evolution, 25(6), 1081–1092.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. PLoS Genetics, 4(12), e1000304, 1-10.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M., White, B. (2010). Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science, 330(6003), 512–514.

Leonardo, T. E., & Muiru, G. T. (2003). Facultative symbionts are associated with host plant specialization in pea aphid populations. Proceedings of the Royal Society B: Biological Sciences, 270(S2), S209–S212.

Legeai, F., Shigenobu, S., Gauthier, J. P., Colbourne, J., Rispe, C., Collin, O. (2010). AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Molecular Biology, 19, 5–12.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078–2079.

McBride, C. (2007). Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences, 104(12), 4996-5001.

Mclean, A. H. C., Van Asch, M., Ferrari, J., & Godfray, H. C. J. (2011). Effects of bacterial secondary symbionts on host plant use in pea aphids. Proceedings of the Royal Society B: Biological Sciences, 278(1706), 760–766.

Moran, N. (1984). The genus *Uroleucon* (Homoptera: Aphididae) in Michigan: key, host records, biological notes, and descriptions of three new species. Journal of the Kansas Entomological Society, 57(4), 596–616.

Moran, N. (1986). Benefits of host plant specificity in *Uroleucon* (Homoptera: Aphididae). Ecology, 67(1),108–115.

Moran, N., Kaplan, M., Gelsey, M., & Murphy, T. (1999). Phylogenetics and evolution of the aphid genus *Uroleucon* based on mitochondrial and nuclear DNA. Systematic Entomology. 24, 85-93.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12(6), 443–451.

Nosil, P. (2002). Transition rates between specialization and generalization in phytophagous insects. Evolution, 56(8), 1701–1706.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587), 332–342.

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. Trends in Ecology & Evolution, 26(4), 160–167.

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. Molecular Ecology, 18(3), 375–402.

Price, D., Karley, A. J., Ashford, D. A., Isaacs, H. V., Pownall, M. E., Wilkinson, H. S. (2007). Molecular characterisation of a candidate gut sucrase in the pea aphid, *Acyrthosiphon pisum*. Insect Biochemistry and Molecular Biology, 37(4), 307–317.

Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F., Simon, J.-C. (2007). Large gene family expansion and variable selective pressures for cathepsin B in aphids. Molecular Biology and Evolution, 25(1), 5–17.

Robinson, A. G. (1985). Annotated list of *Uroleucon* (*Uroleucon*, *Uromelan*, Satula) (Homoptera: Aphididae) of America north of Mexico, with keys and descriptions of new species. The Canadian Entomologist, 117(08), 1029–1054.

Sabri, A., Vandermoten, S., Leroy, P. D., Haubruge, E., Hance, T., Thonart, P. (2013). Proteomic investigation of aphid honeydew reveals an unexpected diversity of proteins. PLoS ONE, 8(9), e74656, 1-10.

Schoonhoven, L. M., van Loon, J. J. A., & Dicke, M. (2005). Insect-Plant Biology. Oxford University Press.

Sexton, J. P., Hangartner, S. B., & Hoffmann, A. A. (2013). Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution, 68(1), 1–15.

Shafer, A. B. A., & Wolf, J. B. W. (2013). Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. Ecology Letters, 16(7), 940–950.

Silva, F. J., Belda, E., & Talens, S. E. (2006). Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. Nucleic Acids Research, 34(20), 6015–6022.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J., & Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution, 66(9), 2723–2738.

Swanson, W. J. (2004). Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics, 168(3), 1457–1465.

Tsuchida, T., Koga, R., Matsumoto, S., & Fukatsu, T. (2011). Interspecific symbiont transfection confers a novel ecological trait to the recipient insect. Biology Letters, 7(2), 245–248.

Van Belleghem, S. M., Roelofs, D., Van Houdt, J., & Hendrickx, F. (2012). *De novo*

transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). PLoS ONE, 7(8), e42605, 1-10.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods, 5(3), 247–252.

Winkler, I., & Mitter, C. (2008). The phylogenetic dimension of insect-plant interactions: a review of recent evidence. Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects, 240–263.

Wright, S. (1943). Isolation by distance. Genetics, 28(2), 114–138.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics & Bioinformatics, 4(4), 259–263.

Complete References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Andrés, J. A. J., Larson, E. L. E., Bogdanowicz, S. M. S., & Harrison, R. G. R. (2013). Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics, 193(2), 501–513.

Baldo, L., Santos, M., & Salzburger, W. (2011). Comparative transcriptomics of eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. Genome Biology and Evolution, 3, 443-455.

Barreto, F. S. F., Moy, G. W. G., & Burton, R. S. R. (2011). Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus*. Molecular Ecology, 20(3), 560–572.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biology, 5(11), 2534-2559.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2010). GenBank. Nucleic Acids Research, *38* (Database issue), D46–D51.

Berenbaum, M. (2002). Postgenomic chemical ecology: from genetic code to ecological interactions. Journal of Chemical Ecology, 28(5), 873–896.

Bernays, E. A., & Chapman, R. F. (1994). Host-plant selection by phytophagous insects. Springer.

Bernays, E. A., & Funk, D. J. (2000). Electrical penetration graph analysis reveals population differentiation of host-plant probing behaviors within the aphid species *Uroleucon ambrosiae*. Entomologia Experimentalis et Applicata, 97(2), 183–191.

Bernays, E., Funk, D., & Moran, N. (2000). Intraspecific differences in olfactory sensilla in relation to diet breadth in *Uroleucon ambrosiae*. Journal of Morphology. 245, 99-109.

Blackman, R. L., & Eastop, V. F. (2008). *Aphids on the World's Herbaceous Plants and Shrubs*. John Wiley & Sons.

Brieuc, M. S. O., & Naish, K. A. (2011). Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources. Molecular Ecology Resources, 11, 172–183.

Brisson, J. A., & Stern, D. L. (2006). The pea aphid, *Acyrthosiphon pisum*: an emerging genomic model system for ecological, developmental and evolutionary studies. BioEssays, 28(7), 747–755.

Broadway, R. M. (1997). Dietary regulation of serine proteinases that are resistant to serine proteinase inhibitors. Journal of Insect Physiology, 43(9), 855–874.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated

alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973.

Carolan, J. C., Caragea, D., Reardon, K. T., Mutti, N. S., Dittmer, N., Pappan, K. (2011). Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): A dual transcriptomic/proteomic approach. Journal of Proteome Research, 10(4), 1505–1518.

Carolan, J. C., Fitzroy, C. I. J., Ashton, P. D., Douglas, A. E., & Wilkinson, T. L. (2009). The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. Proteomics, 9(9), 2457–2467.

Clark, M., Moran, N., Baumann, P., & Wernegreen, J. (2000). Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. Evolution, *54*(2), 517–525.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674–3676.

Cortés, T., Ortiz-Rivas, B., & Martínez-Torres, D. (2010). Identification and characterization of circadian clock genes in the pea aphid *Acyrthosiphon pisum*. Insect Molecular Biology, 19, 123–139.

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? Genetics, 186(1), 41–43.

Dabney A and Storey JD. qvalue: Q-value estimation for false discovery rate control. R package version 1.38.0.

Davey, J. W., & Blaxter, M. L. (2011). RADSeq: next-generation population genetics. Briefings in Functional Genomics, 9(5), 416–423.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, *7*(1), 214.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Wilson, A. (2011). Geneious v5. 4.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, *29*(8), 1969–1973

Dudev, T., & Lim, C. (2003). Principles governing Mg, Ca, and Zn binding and selectivity in proteins. Chemical Reviews, 103(3), 773–788.

Dworkin, I., & Jones, C. D. (2008). Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics, 181(2), 721–736.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460–2461.

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and plants: a study in coevolution. Evolution; International Journal of Organic Evolution, 18(4), 586–608.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. Molecular Ecology, 17(21), 4586–4596.

Falco, M. C., & Silva-Filho, M. C. (2003). Expression of soybean proteinase inhibitors in transgenic sugarcane plants: effects on natural defense against *Diatraea saccharalis*. Plant Physiology and Biochemistry, 41(8), 761–766.

Ferrari, J., Via, S., & Godfray, H. C. J. (2008). Population differentiation and genetic variation in performance on eight hosts in the pea aphid complex. Evolution; International Journal of Organic Evolution, 62(10), 2508–2524.

Fisher, D. B., Wright, J. P., & Mittler, T. E. (1984). Osmoregulation by the aphid *Myzus persicae*: A physiological role for honeydew oligosaccharides. Journal of Insect Physiology, 30(5), 387–393.

Francis, F., Gerkens, P., Harmel, N., Mazzucchelli, G., De Pauw, E., & Haubruge, E. (2006). Proteomics in *Myzus persicae*: effect of aphid host plant switch. Insect Biochemistry and Molecular Biology, 36(3), 219–227.

Frantz, A., Plantegenest, M., Mieuzet, L., & Simon, J. C. (2006). Ecological specialization correlates with genotypic differentiation in sympatric host-populations of the pea aphid. Journal of Evolutionary Biology, 19(2), 392–401.

Funk, D., Helbling, L., & Wernegreen, J. (2000). Intraspecific phylogenetic congruence among multiple symbiont genomes. Proceedings of the Royal Society B: Biological Sciences, 267(1461), 2517-2521.

Funk, D., & Bernays, E. (2001). Geographic variation in host specificity reveals host range evolution in *Uroleucon ambrosiae* aphids. Ecology, 82(3), 726-739.

Funk, D. J., Nosil, P., & Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. Proceedings of the National Academy of Sciences, 103(9), 3209–3213.

Funk, V. A., & Oberprieler, C. (2009). Compositae metatrees: the next generation. In: Funk,Vicki A., (ed.) Systematics, evolution and biogeography of Compositae. International Association for Plant Taxonomy. 747-777.

Funk, V. A. (ed.) (2009). *Systematics, evolution, and biogeography of Compositae*. International Association for Plant Taxonomy.

Futschik, A., & Schlotterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics, 186(1), 207–218.

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. Annual Review of Ecology and Systematics, 19(1), 207–233.

Futuyma, D. J., & Mitter, C. (1996). Insect-plant Interactions: The evolution of component communities. Philosophical Transactions of the Royal Society B: Biological Sciences, *351*(1345), 1361–1366.

Futuyma, D. J., & Agrawal, A. (2009). Macroevolution and the biological diversity of plants and herbivores. Proceedings of the National Academy of Sciences, 106(43), 18054-18061.

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome

Biology, 9(10), 235–235.

Garvin, M., Saitoh, K., & Gharrett, A. (2010). Application of single nucleotide polymorphisms to non‐model species: a technical review. Molecular Ecology Resources. 10(6), 915-934.

Giordanengo, P., Brunissen, L., Rusterucci, C., Vincent, C., Van Bel, A., Dinant, S. (2010). Compatible plant-aphid interactions: How aphids manipulate plant responses. Comptes Rendus Biologies, 333(6-7), 516–523.

Godfray, H. C. J. (2010). The pea aphid genome. Insect Molecular Biology, 19(S2), 1–4.

Gu, X., Zhang, Z., & Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proceedings of the National Academy of Sciences, 102(3), 707–712.

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research, 33(Web Server issue), W557–W559.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494–1512.

Hansen, A., & Moran, N. (2011). Aphid genome expression reveals host–symbiont cooperation in the production of amino acids. Proceedings of the National Academy of Sciences, 108(7), 2849-2854.

Harris, S. E., Munshi-South, J., Obergfell, C., & O'Neill, R. (2013). Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York metropolitan area. PLoS ONE, 8(8), e74938–e74938.

Hawthorne, D., & Via, S. (2001). Genetic linkage of ecological specialization and reproductive isolation in pea aphids. Nature, 412(6850), 904–907.

Heie, O. (1987). Paleontology and phylogeny. In A. K. Minks & P. Harrewijn (Eds.), Aphids: Their Biology, Natural Enemies and Control. Amsterdam: Elsevier.

Heie, O. (1996). The evolutionary history of aphids and a hypothesis on the coevolution of aphids and plants. Bollettino Di Zoologia Agraria E Di Bachicoltura, 28, 149–155.

Hogenhout, S. A., Van der Hoorn, R. A. L., Terauchi, R., & Kamoun, S. (2009). Emerging concepts in effector biology of plant-associated organisms. Molecular Plant-Microbe Interactions, 22(2), 115–122.

Hogenhout, S. A., & Bos, J. I. (2011). Effector proteins that modulate plant–insect interactions. Current Opinion in Plant Biology, 14(4), 422–428.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., & Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Research, 42(Database issue), D897–902.

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

International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biology, *8*(2), e1000313, 1-24.

Janz, N., & Nylin, S. (2008). The oscillation hypothesis of host-plant range and speciation. Specialization, speciation, and radiation: the evolutionary biology of herbivorous insects. University of California Press, Berkeley, 203–215.

Jean, P., & Jean-Christophe, S. (2010). The pea aphid complex as a model of ecological speciation. Ecological Entomology, 35, 119–130.

Jiang, Z. F., Xia, F., Johnson, K. W., Brown, C. D., Bartom, E., Tuteja, J. H. (2013). Comparison of the genome sequences of "*Candidatus Portiera aleyrodidarum*" primary endosymbionts of the whitefly *Bemisia tabaci* B and Q biotypes. Applied and Environmental Microbiology, 79(5), 1757–1759.

Jongsma, M. A., & Bolter, C. (1997). The adaptation of insects to plant protease inhibitors. Journal of Insect Physiology, *43*(10), 885–895.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research, 40(D1), D109–D114.

Karatolos, N., Pauchet, Y., Wilkinson, P., Chauhan, R., Denholm, I., Gorman, K. (2011). Pyrosequencing the transcriptome of the greenhouse whitefly*, Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. BMC Genomics, 12(56), 1-14.

Kelley, S. T., & Farrell, B. D. (1998). Is specialization a dead end? The phylogeny of host use in Dendroctonus bark beetles (Scolytidae). Evolution, 52(6), 1731–1743.

Kim, H., & Lee, S. (2008). A molecular phylogeny of the tribe Aphidini (Insecta: Hemiptera: Aphididae) based on the mitochondrial tRNA/COII, 12S/16S and the nuclear EF1α genes. Systematic Entomology, *33*(4), 711–721.

Kim, H., Lee, S., & Jang, Y. (2011). Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): Diversification, host association, and biogeographic origins. PLoS One, 6(9), e24749, 1-17.

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics, 27(24), 3435–3436.

Konczal, M., Koteja, P., & Stuglik, M. T. (2014). Accuracy of allele frequency estimation using pooled RNA-Seq. Molecular Ecology, 14: 381-392.

Kopp, A., Barmina, O., Hamilton, A. M., Higgins, L., Mcintyre, L. M., & Jones, C. D. (2008). Evolution of gene expression in the *Drosophila* olfactory system. Molecular Biology and Evolution, 25(6), 1081–1092.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. PLoS Genetics, 4(12), e1000304, 1-10.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25.1–

R25.10.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29(6), 1695–1701.

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M., White, B. (2010). Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science, 330(6003), 512–514.

Legeai, F., Shigenobu, S., Gauthier, J. P., Colbourne, J., Rispe, C., Collin, O. (2010). AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Molecular Biology, 19, 5–12.

Leonardo, T. E., & Muiru, G. T. (2003). Facultative symbionts are associated with host plant specialization in pea aphid populations. Proceedings of the Royal Society B: Biological Sciences, 270(S2), S209–S212.

Legeai, F., Shigenobu, S., Gauthier, J. P., Colbourne, J., Rispe, C., Collin, O. (2010). AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Molecular Biology, 19, 5–12.

Lespinet, O. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Research, 12(7), 1048–1059.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078–2079.

McBride, C. (2007). Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences, 104(12), 4996-5001.

Mclean, A. H. C., Van Asch, M., Ferrari, J., & Godfray, H. C. J. (2011). Effects of bacterial secondary symbionts on host plant use in pea aphids. Proceedings of the Royal Society B: Biological Sciences, 278(1706), 760–766.

Milne, R.I. (2006). Northern Hemisphere plant disjunctions: a window on Tertiary land bridges and climate change? Annals of Botany, *98*(3), 465–472.

Mitter, C. C., Farrell, B. B., & Futuyma, D. J. D. (1991). Phylogenetic studies of insect-plant interactions: Insights into the genesis of diversity. Trends in Ecology & Evolution, 6(9), 290–293.

Mitter, C., Farrell, B., & Wiegmann, B. (1988). The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? American Naturalist, 132(1), 107.

Moran, N. (1984). The genus *Uroleucon* (Homoptera: Aphididae) in Michigan: key, host records, biological notes, and descriptions of three new species. Journal of the Kansas Entomological Society, 57(4), 596–616.

Moran, N. (1986). Benefits of host plant specificity in *Uroleucon* (Homoptera: Aphididae). Ecology, 67(1),108–115.

Moran, N., Kaplan, M., Gelsey, M., & Murphy, T. (1999). Phylogenetics and evolution of the aphid genus *Uroleucon* based on mitochondrial and nuclear DNA. Systematic Entomology. 24, 85-93.

Mutti, N. S., Louis, J., Pappan, L. K., Pappan, K., Begum, K., Chen, M.-S. (2008). A protein from the salivary glands of the pea aphid, *Acyrthosiphon pisum*, is essential in feeding on a host plant. Proceedings of the National Academy of Sciences, 105(29), 9965–9969.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12(6), 443–451.

Nosil, P. (2002). Transition rates between specialization and generalization in phytophagous insects. Evolution, 56(8), 1701–1706.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587), 332–342.

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. Trends in Ecology & Evolution, 26(4), 160–167.

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. Molecular Ecology, 18(3), 375–402.

Nováková, E., Hypša, V., Klein, J., Foottit, R. G., von Dohlen, C. D., & Moran, N. A. (2013). Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola. Molecular Phylogenetics and Evolution, 68(1), 42–54.

Ohno, S. (1970). Evolution by gene duplication. New York: Springer-Verlag.

Olive, A. T. (1965). A new subgenus and two new species of *Dactynotus* (Homoptera: Aphididae). Annals of the Entomological Society of America, 58(3), 284-289.

Oliver, K. M., Degnan, P. H., Burke, G. R., & Moran, N. A. (2010). Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. Annual Review of Entomology, *55*, 247–266.

Ollivier, M., Legeai, F., & Rispe, C. (2010). Comparative analysis of the *Acyrthosiphon pisum* genome and expressed sequence tag-based gene sets from other aphid species. Insect Molecular Biology, 19, 33–45.

Ortiz-Rivas, B., & Martínez-Torres, D. (2010). Combination of molecular data support the existence of three main lineages in the phylogeny of aphids (Hemiptera: Aphididae) and the basal position of the subfamily Lachninae. Molecular Phylogenetics and Evolution, 55(1), 305–317.

Pagel, M., & Meade, A. (2007). BayesTraits. Software Distributed by the Author. http://www.evolution.reading.ac.uk/BayesTraits.html

Pagel, M., Meade, A., & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*. 53(5), 673-684.

Papasotiropoulos, V., & Tsiamis, G. (2013). A molecular phylogenetic study of aphids (Hemiptera: Aphididae) based on mitochondrial DNA sequence analysis. Journal of Biological Research – Thessaloniki, 20, 1-13.

Peccoud, J., Ollivier, A., Plantegenest, M., & Simon, J. (2009). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proceedings of the National Academy of Sciences, 106(18), 7495-7500.

Peccoud, J., Simon, J. von Dohlen, C., Coeur d'acier, A., Plantegenest, M., Vanlerberghe-Masutti, F., & Jousselin, E. (2010). Evolutionary history of aphid-plant associations and their role in aphid diversification. Comptes Rendus Biologies, 333(6-7), 474–487.

Pelser, P. B., & Watson, L. E. (2009). Introduction to Asteroideae. In: Funk,Vicki A., (ed.) Systematics, evolution and biogeography of Compositae. International Association for Plant Taxonomy, 495-502.

Price, D., Karley, A. J., Ashford, D. A., Isaacs, H. V., Pownall, M. E., Wilkinson, H. S. (2007). Molecular characterisation of a candidate gut sucrase in the pea aphid, *Acyrthosiphon pisum*. Insect Biochemistry and Molecular Biology, 37(4), 307–317.

Ramsey, J. S., Macdonald, S. J., Jander, G., Nakabachi, A., Thomas, G. H., & Douglas, A. E. (2010). Genomic evidence for complementary purine metabolism in the pea aphid, *Acyrthosiphon pisum*, and its symbiotic bacterium *Buchnera aphidicola*. Insect Molecular Biology, *19*, 241–248.

Ramsey, J., Rider, D., Walsh, T., De Vos, M., Gordon, K., Ponnala, L. (2010). Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae.* Insect Molecular Biology, 19, 155–164.

Riddell, C. E., Sumner, S., Adams, S., & Mallon, E. B. (2011). Pathways to immunity: temporal dynamics of the bumblebee (*Bombus terrestris*) immune response against a trypanosomal gut parasite. Insect Molecular Biology, *20*(4), 529–540.

Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F., Simon, J.-C. (2007). Large gene family expansion and variable selective pressures for cathepsin B in aphids. Molecular Biology and Evolution, 25(1), 5–17.

Robinson, A. G. (1985). Annotated list of *Uroleucon* (*Uroleucon*, *Uromelan*, Satula) (Homoptera: Aphididae) of America north of Mexico, with keys and descriptions of new species. The Canadian Entomologist, 117(08), 1029–1054.

Robinson, A. G. (1986). Annotated list of *Uroleucon* (*Lambersius*) (Homoptera: Aphididae) of America north of Mexico, with a key and descriptions of new species. The Canadian Entomologist, 118(6), 559–576.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), R25.1 – R25.10.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140.

Sabri, A., Vandermoten, S., Leroy, P. D., Haubruge, E., Hance, T., Thonart, P. (2013). Proteomic investigation of aphid honeydew reveals an unexpected diversity of proteins. PLoS ONE, 8(9), e74656, 1-10.

Schoonhoven, L. M., van Loon, J. J. A., & Dicke, M. (2005). Insect-Plant Biology. Oxford University Press.

Sexton, J. P., Hangartner, S. B., & Hoffmann, A. A. (2013). Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution, 68(1), 1–15.

Shafer, A. B. A., & Wolf, J. B. W. (2013). Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. Ecology Letters, 16(7), 940–950.

Shigenobu, S., Bickel, R. D., Brisson, J. A., Butts, T., Chang, C. C., Christiaens, O. (2010). Comprehensive survey of developmental genes in the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes. Insect Molecular Biology, 19, 47–62.

Silva, F. J., Belda, E., & Talens, S. E. (2006). Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. Nucleic Acids Research, 34(20), 6015–6022.

Simon, J. C., Carre, S., Boutin, M., Prunier-Leterme, N., Sabater-Munoz, B., Latorre, A., Bournoville, R. (2003). Host-based divergence in populations of the pea aphid: insights from nuclear markers and the prevalence of facultative symbionts. Proceedings of the Royal Society B: Biological Sciences, 270(1525), 1703–1712.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19(6), 1117–1123.

Smadja, C. M., Canbäck, B., Vitalis, R., Gautier, M., Ferrari, J., Zhou, J.-J., & Butlin, R. K. (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution, 66(9), 2723–2738.

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M. (2008). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. Botanical Journal of the Linnean Society, 133(4), 381–461.

Soltis, D. E., Soltis, P. S., Endress, P. K., Chase, M. W., Soltis, D. E., Soltis, P. S. (2005). Phylogeny and evolution of angiosperms. Sinauer Associates Incorporated.

Soltis, P. S., Solits, D. E., & Chase, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature, 402(6760), 402–404.

Swanson, W. J. (2004). Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics, 168(3), 1457–1465.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), 2688–2690.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 479–498.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Science, 100, 9440-9445.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution, 28(10), 2731–2739.

Thompson, G. A. (2006). Transcriptomics and functional genomics of plant defence induction by

phloem-feeding insects. Journal of Experimental Botany, 57(4), 755–766.

Tsuchida, T., Koga, R., Matsumoto, S., & Fukatsu, T. (2011). Interspecific symbiont transfection confers a novel ecological trait to the recipient insect. Biology Letters, 7(2), 245–248.

Van Belleghem, S. M., Roelofs, D., Van Houdt, J., & Hendrickx, F. (2012). *De novo* transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). PLoS ONE, 7(8), e42605, 1-10.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods, 5(3), 247–252.

Via, S., & Hawthorne, D. (2002). The genetic architecture of ecological specialization: correlated gene effects on host use and habitat choice in pea aphids. American Naturalist, 159(3), 76–88.

von Dohlen, C. (2000). Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. Biological Journal of the Linnean Society, 71(4), 689–717.

von Dohlen, C., Rowe, C., & Heie, O. (2006). A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. Molecular Phylogenetics and Evolution, 38(2), 316–329.

Vieira, F., & Rozas, J. (2011). Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biology and Evolution, 3, 476-490.

Wagner, A. (2000). Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. Proceedings of the National Academy of Sciences, 97(12), 6579–6584.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D., Brockington, S. F., Alexandre, R. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. Proceedings of the National Academy of Sciences, 106(10), 3853–3858.

Wiens, J. J. J., & Tiu, J. J. (2012). Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS ONE, *7*(8), e42925.

Wiens, J. J., & Morrill, M. C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Systematic Biology, 60(5), 719–731.

Will, T. (2006). Physical and chemical interactions between aphids and plants. Journal of Experimental Botany, 57(4), 729–737.

Will, T., Tjallingii, W. F., Thönnessen, A., & van Bel, A. J. E. (2007). Molecular sabotage of plant defense by aphid saliva. Proceedings of the National Academy of Sciences of the United States of America, 104(25), 10536–10541.

Winkler, I., & Mitter, C. (2008). The phylogenetic dimension of insect-plant interactions: a review of recent evidence. Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects, 240–263.

Wright, S. (1943). Isolation by distance. Genetics, 28(2), 114–138.

Yoon, O. K., & Brem, R. B. (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA, 1–13.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6), 292–298.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics & Bioinformatics, 4(4), 259–263.

Chapter 2 figure captions

Figure 1. Best supported cladograms for Macrosiphini-dataset
Tree topologies based on combined analysis of six loci (Macrosiphini-dataset): three mitochondrial (COI, COII, tRNA-val), one nuclear (elongation factor 1-a) and two *Buchnera* loci (trp, groEL). a) consensus maximum parsimony (MP) tree, b) most likely maximum likelihood (ML) tree, c) Bayesian maximum clade credibility tree. Bootstrap support values are given for each node of the ML and MP trees, and posterior probability values are given for the Bayesian tree. All support values are scaled to range from 0 to 1. Nodes with less than 0.5 support are collapsed. Several genera are represented by multiple species—each such genus is collapsed if monophyletic. The tribal designation for each species is given on the right margin.

Figure 2. Combined Macrosiphini-dataset tree with tribal classifications
Combined Bayesian tree based on Macrosiphini-dataset indicating tribal classifications. The Bayesian posterior probability appears above each branch. Below each branch, from left to right are ML and MP bootstrap values. Branches without at least two support values over 50% are indicated in gray. Support values in brackets indicate support for an alternative topology.

Figure 3. *Uroleucon* tree with sub-generic classifications
Cladogram for *Uroleucon* species included in *Uroleucon*-dataset. The topology is based on the maximum clade credibility tree based on multilocus Bayesian inference. Branches are shaded according to their posterior probability value. Many branches are poorly supported, especially among basal lineages, and among species closely related to *U. ambrosiae.* Sub-generic designations are given for each species; none are monophyletic, replicating results found previously (Moran *et al.*, 1999).

Figure 4. Phylogenetic mapping of host-use traits in *Uroleucon*
Maximum clade credibility tree based on multi-locus Bayesian analysis of *Uroleucon*-dataset. Branch lengths are based on a relaxed molecular clock model and reflect divergence times in millions of years as indicated by the scale, based on a calibration date of 48 MYA for the common ancestor of Macrosiphini. Divergence times appear in green next to relevant nodes. Branches are shaded according to their posterior probability value. For each species, host sub-families are indicated by shading of taxon names, host tribes are given in the right margin, and bar graphs reflect the number of genera reliably recorded as hosts. Specialization on a single genus is given a value of 1. Host-use data is based on published descriptions (Blackman & Eastop, 2008; Heie, 1996; Robinson, 1985; Robinson, 1986).

Figure 5. Ancestral state reconstruction of biogeographic origin in *Uroleucon*
Maximum clade credibility tree based on multi-locus Bayesian analysis of *Uroleucon*-dataset. See caption for Fig. 5 for details on the inference method and branch length estimation. Biogeographic origin is given for each taxon based on published records (Blackman & Eastop, 2008). Reconstructed ancestral biogeographic range states are presented as pie charts according to the relative probabilities at ancestral nodes of interest. Relative probabilities are based on classification of biogeographic range into one of four states, as follows: *Uroleucon* ancestor (55% N. America, 21% Europe/N. Africa/central Asia, 19% Asia/E. Asia, 5% Holarctic); North American clade (97%, 1%, 1%, 0%), *Ua* clade (100%, 0%, 0%, 0%).

Figure 6. Ancestral state reconstruction of host use of Aphidinae tribes
Ancestral host-use states are classified as asterid (green), rosid (blue), generalist (gray) and other (red), including three genera using Poaceae, which diversified along with Rosids and Asterids, coincident with Aphidinae diversification (von Dohlen, 2000; Heie, 1996). Reconstructed ancestral host use traits are presented as pie charts according to the relative probabilities at ancestral nodes of interest ("U+A clade" is the smallest grouping containing the common ancestor of *Uroleucon* and *Acyrthosiphon)*. Values for relative probabilities are as follows: Aphidinae clade (76% Rosid, 1% Asterid, 4% generalist, 19% other); Macrosiphini clade (76%, 1%, 3%, 21%); U+A clade (71%, 5%, 0%, 23%); *Uroleucon* clade (99% Asterid). The topology is derived from Bayesian analysis of the Macrosiphini-dataset (69 taxa). Genera represented by more than one species (number of taxa indicated in brackets) are collapsed. All nodes have posterior probability >50; branches are shaded according to their posterior probability, and are gray if no more than one support value is >50. ML and MP bootstrap values are given below each branch, from left to right. Terminal taxa are colored according to host-use states. Divergence dates (boxed numbers) are from Kim *et al.* (2011).

Figure 7. Ancestral state reconstruction of host-alternation of Aphidinae tribes
Ancestral host-use states are classified as alternating (blue) or simple monoecious (white). While some aphid species are monoecious on woody hosts, all monoecious species in this analysis obligately feed on herbaceous hosts. Species that represent genera with both host-alternating and monoecious traits are classified as host-alternating. States are coded based on published descriptions (Blackman & Eastop, 2008). Reconstructed ancestral host-alternation traits are presented as pie charts. Values for relative probabilities are as follows: Aphidinae clade (64% simple, 36% alternating), Macrosiphini clade (71% simple, 29% alternating), U+A clade (72% simple, 28% alternating), *Ua* clade (99% simple, 1% alternating). The tree topology and support values are as in Figure 4.

Figure 1b

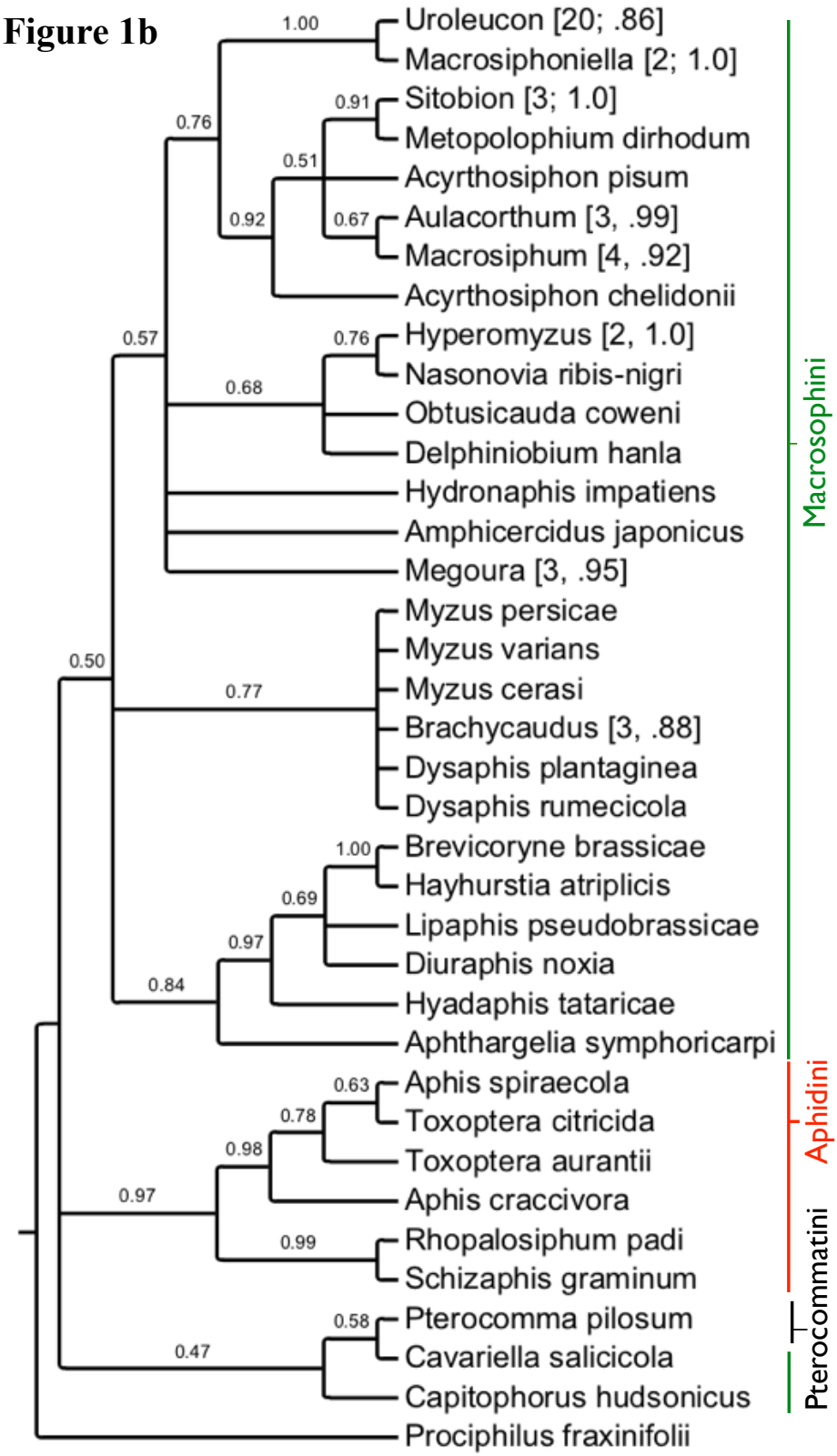

Uroleucon [20; .86]
Macrosiphoniella [2; 1.0]
Sitobion [3; 1.0]
Metopolophium dirhodum
Acyrthosiphon pisum
Aulacorthum [3, .99]
Macrosiphum [4, .92]
Acyrthosiphon chelidonii
Hyperomyzus [2, 1.0]
Nasonovia ribis-nigri
Obtusicauda coweni
Delphiniobium hanla
Hydronaphis impatiens
Amphicercidus japonicus
Megoura [3, .95]
Myzus persicae
Myzus varians
Myzus cerasi
Brachycaudus [3, .88]
Dysaphis plantaginea
Dysaphis rumecicola
Brevicoryne brassicae
Hayhurstia atriplicis
Lipaphis pseudobrassicae
Diuraphis noxia
Hyadaphis tataricae
Aphthargelia symphoricarpi
Aphis spiraecola
Toxoptera citricida
Toxoptera aurantii
Aphis craccivora
Rhopalosiphum padi
Schizaphis graminum
Pterocomma pilosum
Cavariella salicicola
Capitophorus hudsonicus
Prociphilus fraxinifolii
1.00
0.91
0.76
0.51
0.92
0.67
0.57
0.76
0.68
0.50
0.77
1.00
0.69
0.97
0.84
0.63
0.78
0.98
0.97
0.99
0.58
0.47
Macrosophini
Aphidini
Pterocommatini

Figure 1a

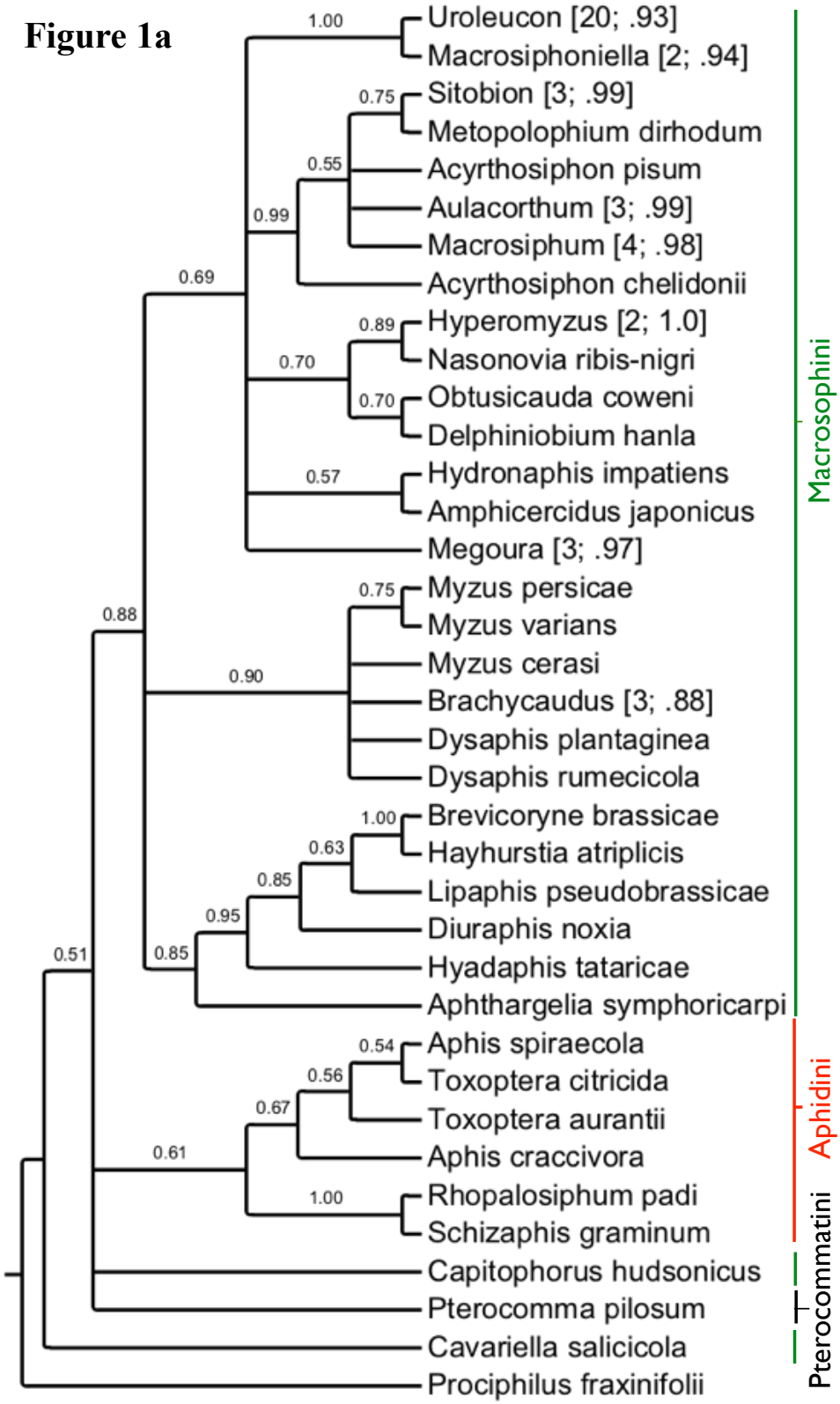

Uroleucon [20; .93]
Macrosiphoniella [2; .94]
Sitobion [3; .99]
Metopolophium dirhodum
Acyrthosiphon pisum
Aulacorthum [3; .99]
Macrosiphum [4; .98]
Acyrthosiphon chelidonii
Hyperomyzus [2; 1.0]
Nasonovia ribis-nigri
Obtusicauda coweni
Delphiniobium hanla
Hydronaphis impatiens
Amphicercidus japonicus
Megoura [3; .97]
Myzus persicae
Myzus varians
Myzus cerasi
Brachycaudus [3; .88]
Dysaphis plantaginea
Dysaphis rumecicola
Brevicoryne brassicae
Hayhurstia atriplicis
Lipaphis pseudobrassicae
Diuraphis noxia
Hyadaphis tataricae
Aphthargelia symphoricarpi
Aphis spiraecola
Toxoptera citricida
Toxoptera aurantii
Aphis craccivora
Rhopalosiphum padi
Schizaphis graminum
Capitophorus hudsonicus
Pterocomma pilosum
Cavariella salicicola
Prociphilus fraxinifolii
1.00
0.75
0.55
0.99
0.69
0.89
0.70
0.70
0.57
0.75
0.88
0.90
1.00
0.63
0.85
0.95
0.51
0.85
0.54
0.56
0.67
0.61
1.00
Macrosophini
Aphidini
Pterocommatini

Figure 1c

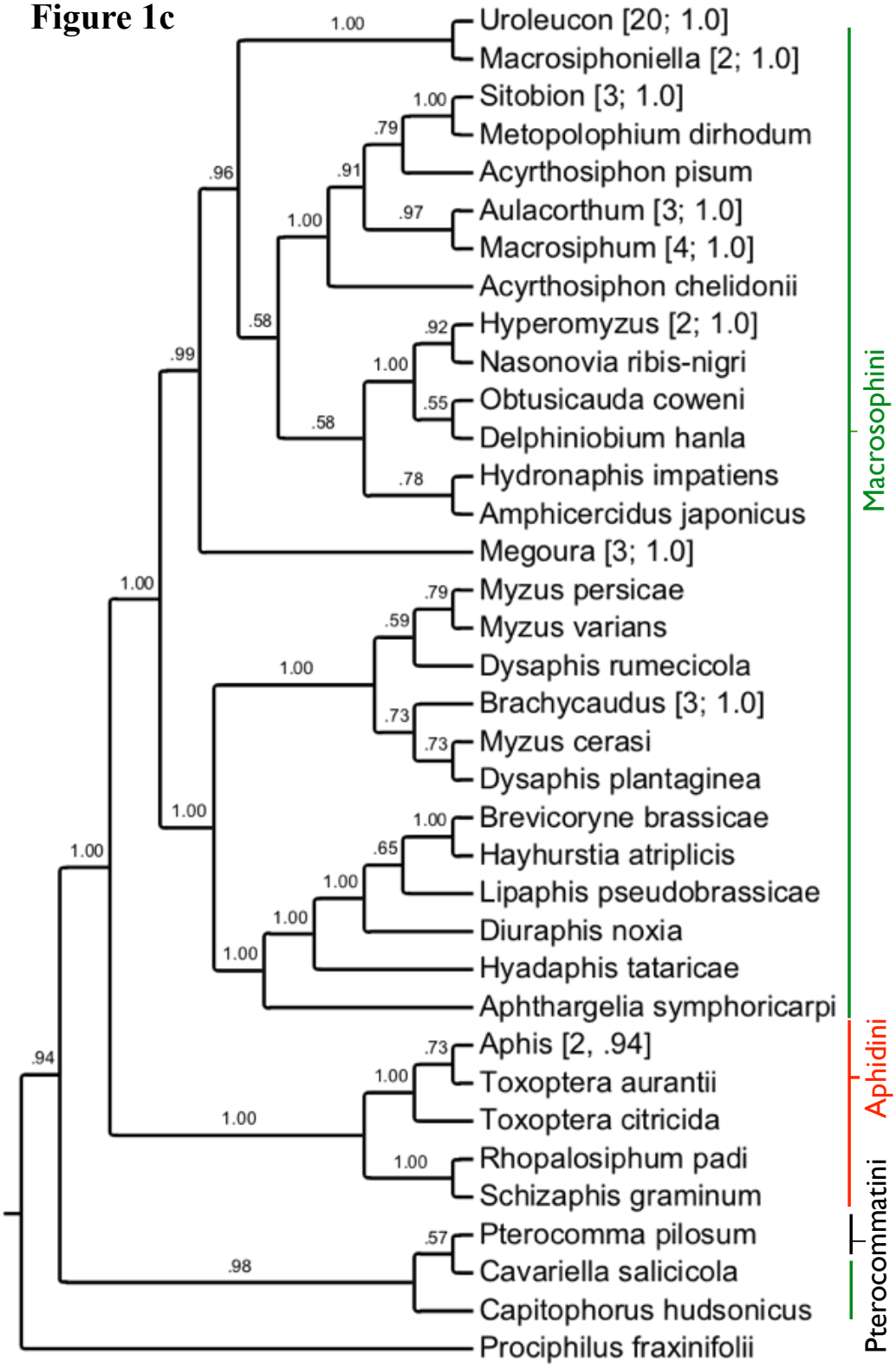

Uroleucon [20; 1.0]
Macrosiphoniella [2; 1.0]
Sitobion [3; 1.0]
Metopolophium dirhodum
Acyrthosiphon pisum
Aulacorthum [3; 1.0]
Macrosiphum [4; 1.0]
Acyrthosiphon chelidonii
Hyperomyzus [2; 1.0]
Nasonovia ribis-nigri
Obtusicauda coweni
Delphiniobium hanla
Hydronaphis impatiens
Amphicercidus japonicus
Megoura [3; 1.0]
Myzus persicae
Myzus varians
Dysaphis rumecicola
Brachycaudus [3; 1.0]
Myzus cerasi
Dysaphis plantaginea
Brevicoryne brassicae
Hayhurstia atriplicis
Lipaphis pseudobrassicae
Diuraphis noxia
Hyadaphis tataricae
Aphthargelia symphoricarpi
Aphis [2, .94]
Toxoptera aurantii
Toxoptera citricida
Rhopalosiphum padi
Schizaphis graminum
Pterocomma pilosum
Cavariella salicicola
Capitophorus hudsonicus
Prociphilus fraxinifolii
Macrosophini
Aphidini
Pterocommatini

Figure 2

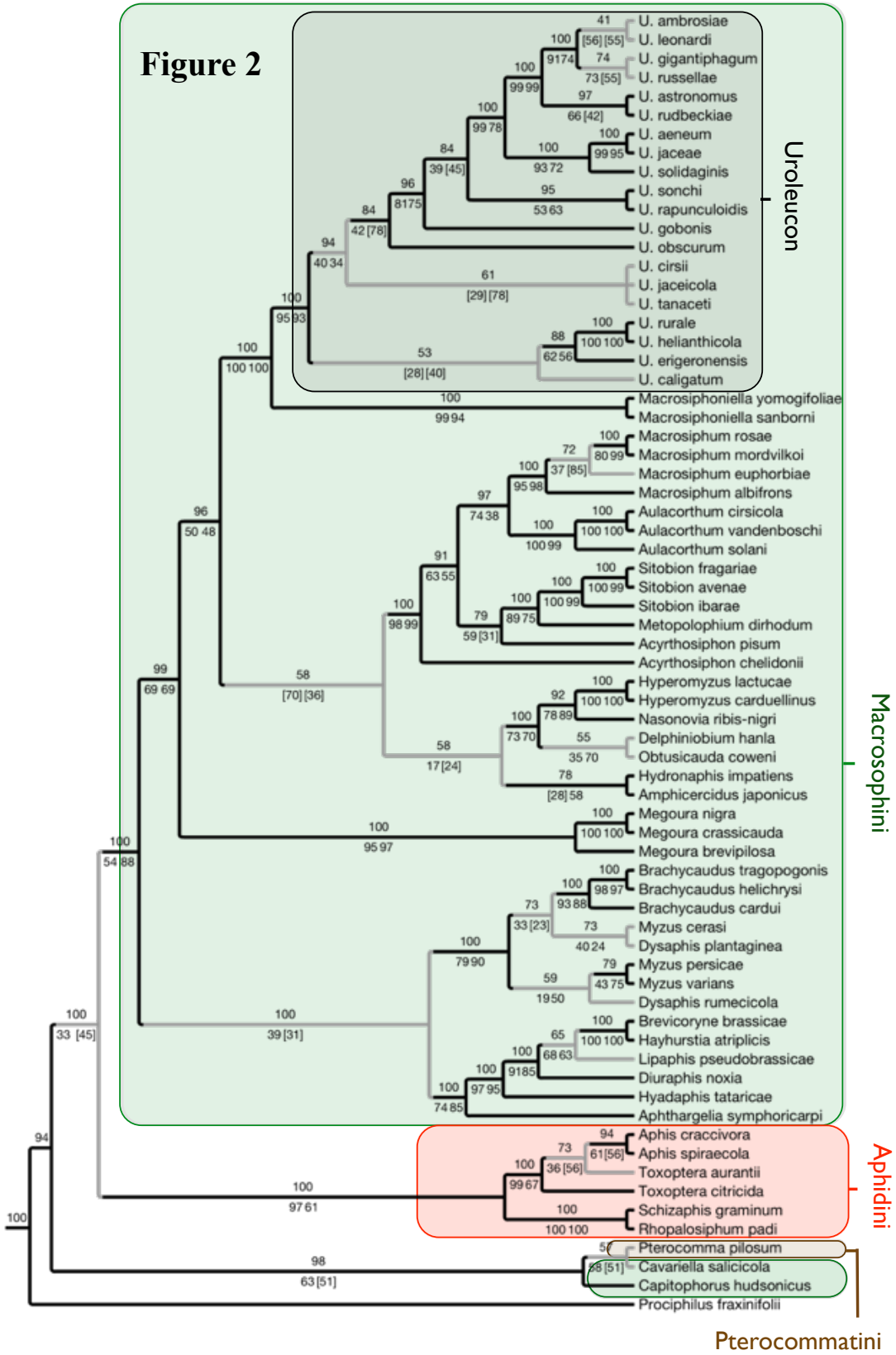

U. ambrosiae
U. leonardi
U. gigantiphagum
U. russellae
U. astronomus
U. rudbeckiae
U. aeneum
U. jaceae
U. solidaginis
U. sonchi
U. rapunculoidis
U. gobonis
U. obscurum
U. cirsii
U. jaceicola
U. tanaceti
U. rurale
U. helianthicola
U. erigeronensis
U. caligatum
Uroleucon
Macrosiphoniella yomogifoliae
Macrosiphoniella sanborni
Macrosiphum rosae
Macrosiphum mordvilkoi
Macrosiphum euphorbiae
Macrosiphum albifrons
Aulacorthum cirsicola
Aulacorthum vandenboschi
Aulacorthum solani
Sitobion fragariae
Sitobion avenae
Sitobion ibarae
Metopolophium dirhodum
Acyrthosiphon pisum
Acyrthosiphon chelidonii
Hyperomyzus lactucae
Hyperomyzus carduellinus
Nasonovia ribis-nigri
Delphiniobium hanla
Obtusicauda coweni
Hydronaphis impatiens
Amphicercidus japonicus
Megoura nigra
Megoura crassicauda
Megoura brevipilosa
Brachycaudus tragopogonis
Brachycaudus helichrysi
Brachycaudus cardui
Myzus cerasi
Dysaphis plantaginea
Myzus persicae
Myzus varians
Dysaphis rumecicola
Brevicoryne brassicae
Hayhurstia atriplicis
Lipaphis pseudobrassicae
Diuraphis noxia
Hyadaphis tataricae
Aphthargelia symphoricarpi
Macrosophini
Aphis craccivora
Aphis spiraecola
Toxoptera aurantii
Toxoptera citricida
Schizaphis graminum
Rhopalosiphum padi
Aphidini
Pterocomma pilosum
Cavariella salicicola
Capitophorus hudsonicus
Prociphilus fraxinifolii
Pterocommatini

Figure 3

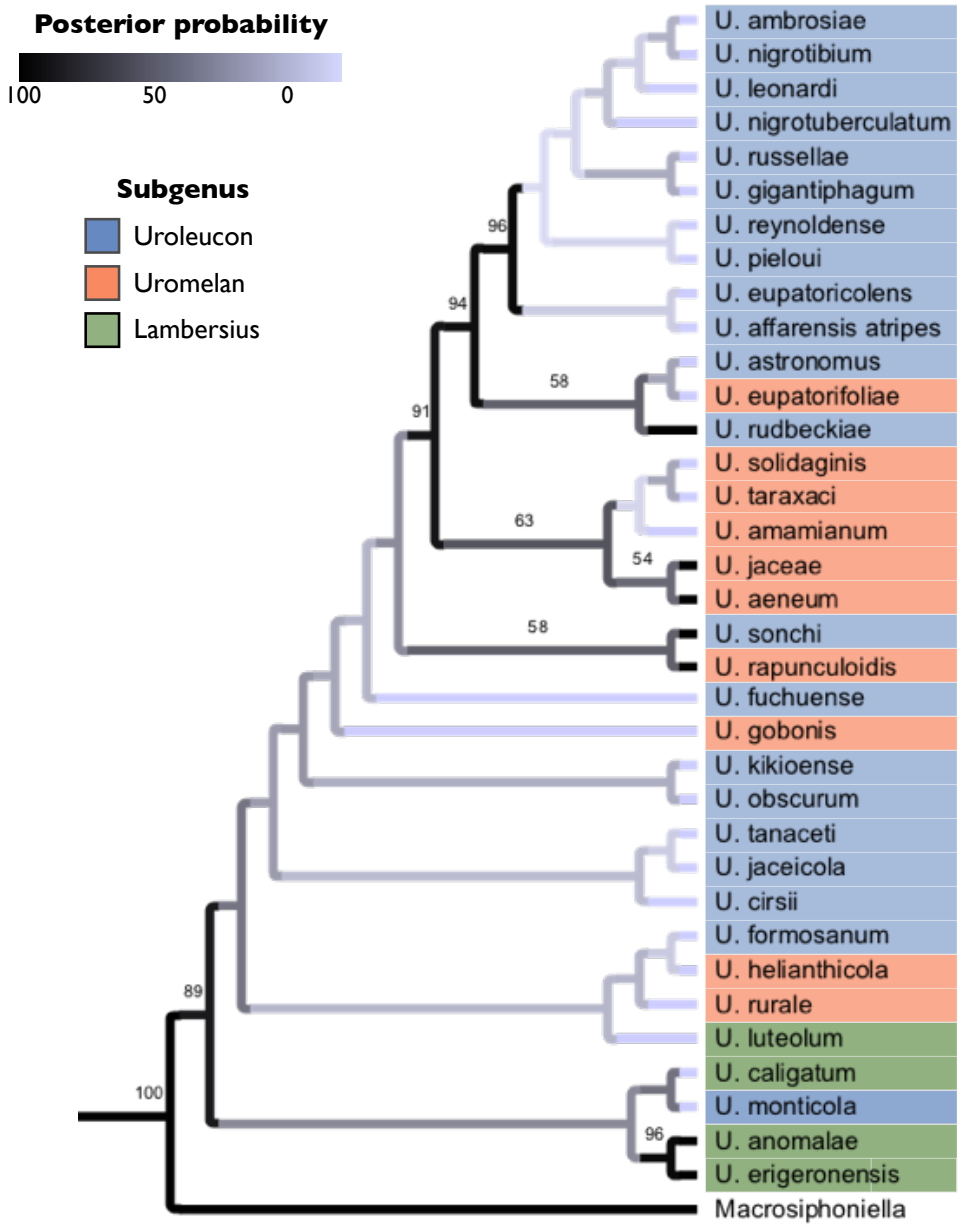

Posterior probability
100
50
0
Subgenus
Uroleucon
Uromelan
Lambersius
U. ambrosiae
U. nigrotibium
U. leonardi
U. nigrotuberculatum
U. russellae
U. gigantiphagum
U. reynoldense
U. pieloui
U. eupatoricolens
U. affarensis atripes
U. astronomus
U. eupatorifoliae
U. rudbeckiae
U. solidaginis
U. taraxaci
U. amamianum
U. jaceae
U. aeneum
U. sonchi
U. rapunculoidis
U. fuchuense
U. gobonis
U. kikioense
U. obscurum
U. tanaceti
U. jaceicola
U. cirsii
U. formosanum
U. helianthicola
U. rurale
U. luteolum
U. caligatum
U. monticola
U. anomalae
U. erigeronensis
Macrosiphoniella
96
94
58
91
63
54
58
89
100
96

Figure 4

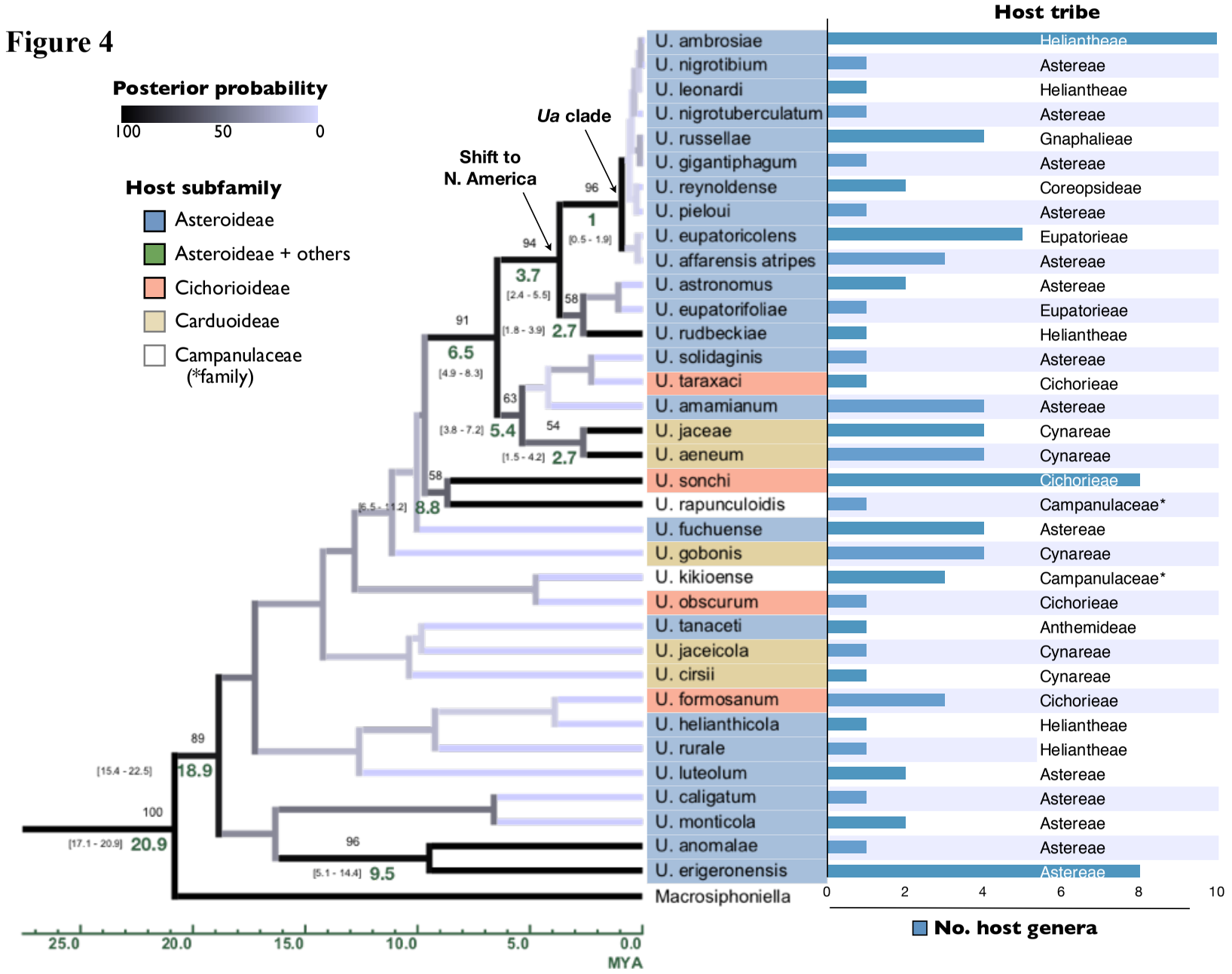

Posterior probability
100
50
0
Host subfamily
Asteroideae
Asteroideae + others
Cichorioideae
Carduoideae
Campanulaceae
(*family)
Ua clade
Shift to
N. America
96
1
[0.5 - 1.9]
94
3.7
[2.4 - 5.5]
58
91
[1.8 - 3.9] 2.7
6.5
[4.9 - 8.3]
63
[3.8 - 7.2] 5.4
54
[1.5 - 4.2] 2.7
58
[6.5 - 11.2] 8.8
89
[15.4 - 22.5] 18.9
100
[17.1 - 20.9] 20.9
96
[5.1 - 14.4] 9.5
25.0
20.0
15.0
10.0
5.0
0.0
MYA
U. ambrosiae
U. nigrotibium
U. leonardi
U. nigrotuberculatum
U. russellae
U. gigantiphagum
U. reynoldense
U. pieloui
U. eupatoricolens
U. affarensis atripes
U. astronomus
U. eupatorifoliae
U. rudbeckiae
U. solidaginis
U. taraxaci
U. amamianum
U. jaceae
U. aeneum
U. sonchi
U. rapunculoidis
U. fuchuense
U. gobonis
U. kikioense
U. obscurum
U. tanaceti
U. jaceicola
U. cirsii
U. formosanum
U. helianthicola
U. rurale
U. luteolum
U. caligatum
U. monticola
U. anomalae
U. erigeronensis
Macrosiphoniella
Host tribe
Heliantheae
Astereae
Heliantheae
Astereae
Gnaphalieae
Astereae
Coreopsideae
Astereae
Eupatorieae
Astereae
Astereae
Eupatorieae
Heliantheae
Astereae
Cichorieae
Astereae
Cynareae
Cynareae
Cichorieae
Campanulaceae*
Astereae
Cynareae
Campanulaceae*
Cichorieae
Anthemideae
Cynareae
Cynareae
Cichorieae
Heliantheae
Heliantheae
Astereae
Astereae
Astereae
Astereae
0
2
4
6
8
10
No. host genera

Figure 5

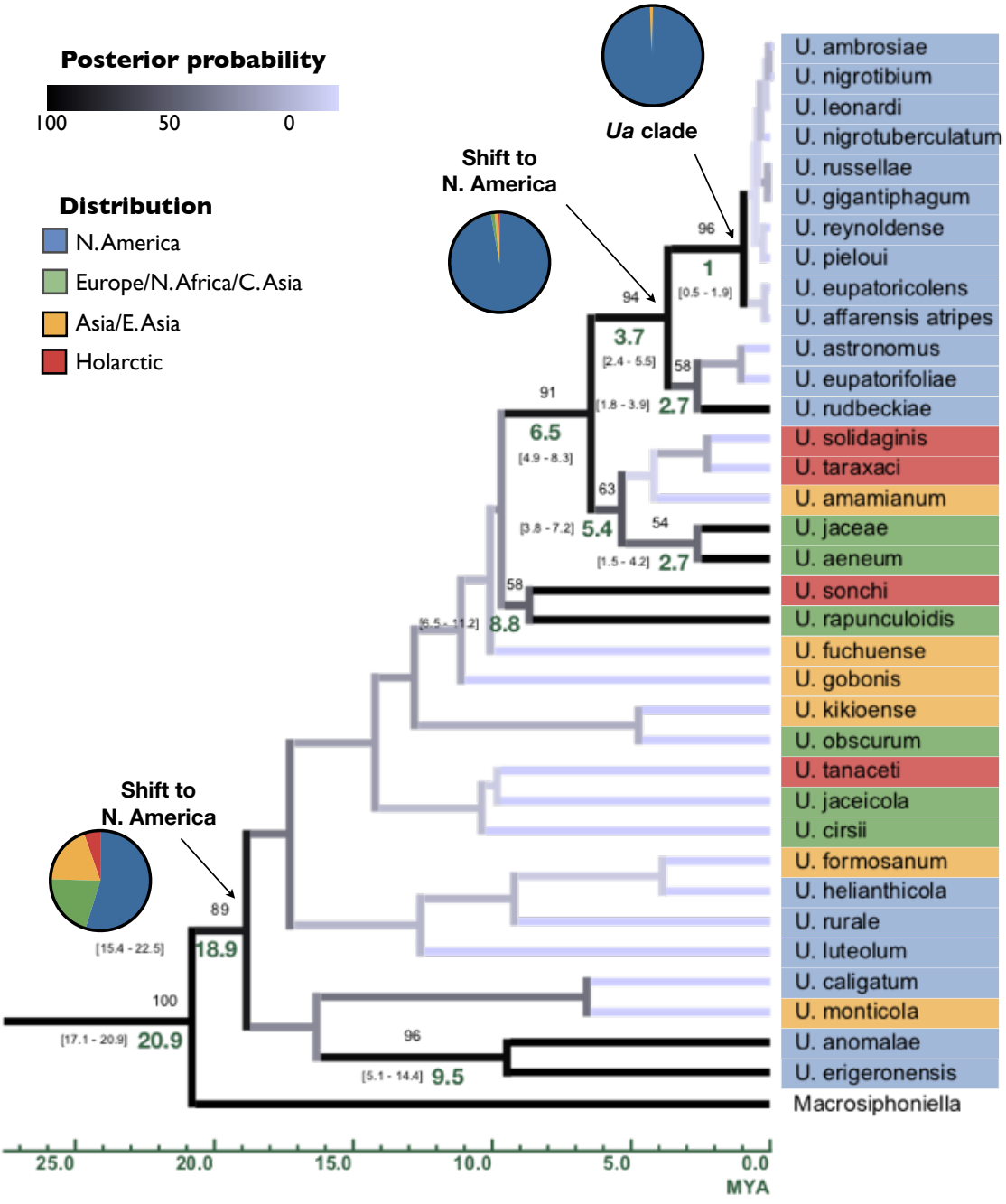
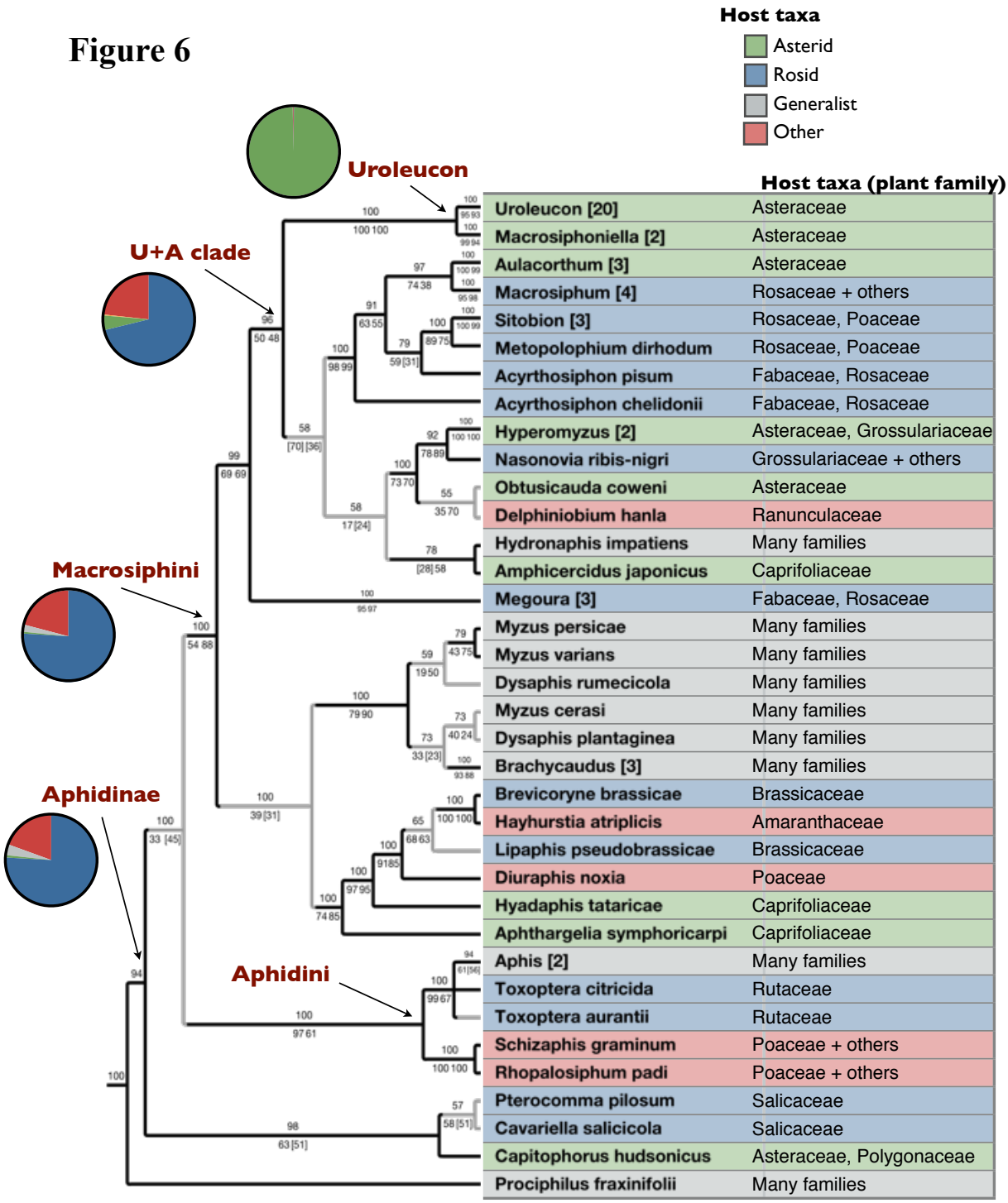
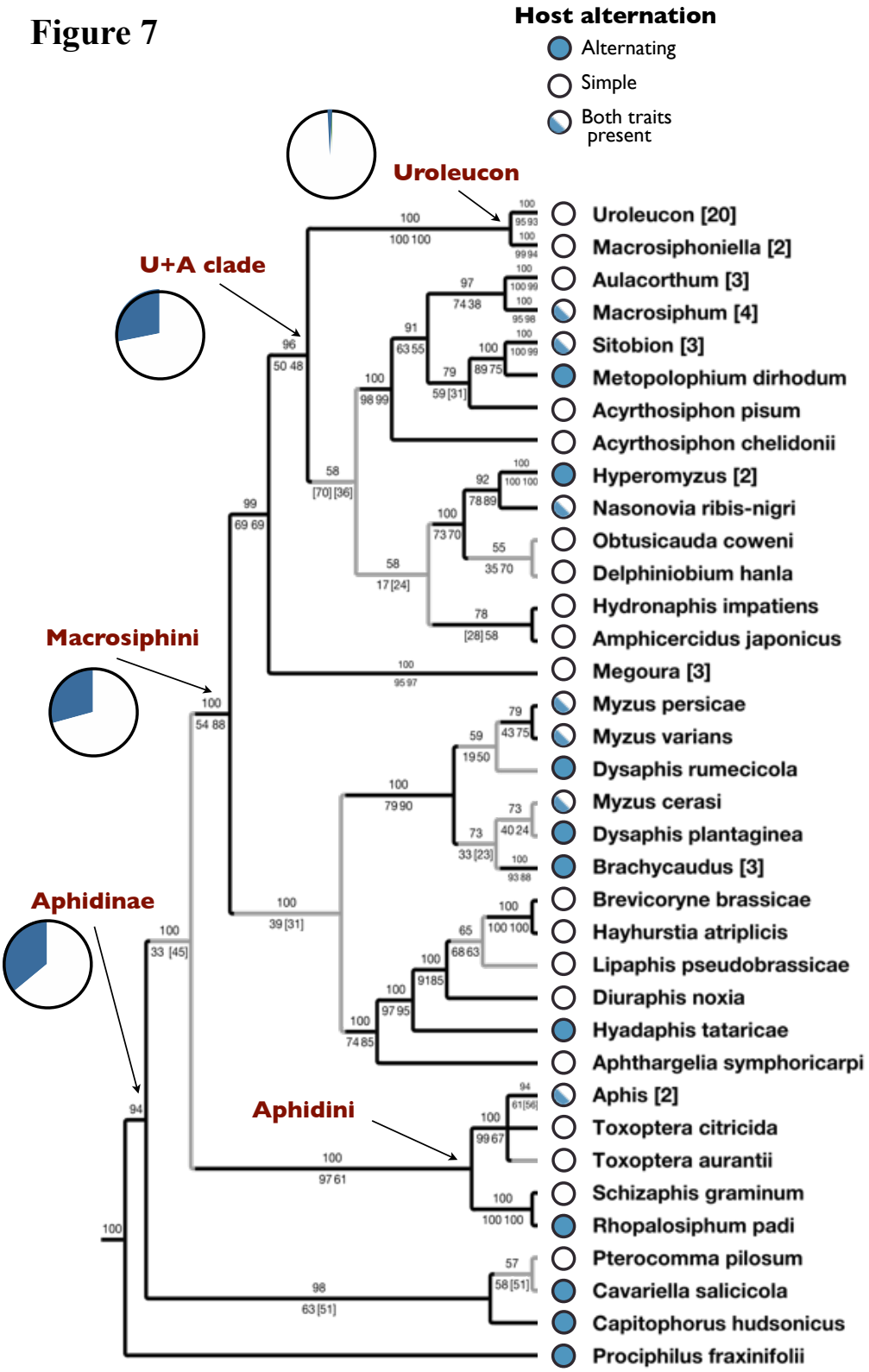

Posterior probability
100
50
0
Distribution
N. America
Europe/N. Africa/C. Asia
Asia/E. Asia
Holarctic
Shift to
N. America
Ua clade
Shift to
N. America
U. ambrosiae
U. nigrotibium
U. leonardi
U. nigrotuberculatum
U. russellae
U. gigantiphagum
U. reynoldense
U. pieloui
U. eupatoricolens
U. affarensis atripes
U. astronomus
U. eupatorifoliae
U. rudbeckiae
U. solidaginis
U. taraxaci
U. amamianum
U. jaceae
U. aeneum
U. sonchi
U. rapunculoidis
U. fuchuense
U. gobonis
U. kikioense
U. obscurum
U. tanaceti
U. jaceicola
U. cirsii
U. formosanum
U. helianthicola
U. rurale
U. luteolum
U. caligatum
U. monticola
U. anomalae
U. erigeronensis
Macrosiphoniella
96
1
[0.5 - 1.9]
94
3.7
[2.4 - 5.5]
58
91
[1.8 - 3.9] 2.7
6.5
[4.9 - 8.3]
63
[3.8 - 7.2] 5.4
54
[1.5 - 4.2] 2.7
58
[6.5 - 11.2] 8.8
89
[15.4 - 22.5]
18.9
100
[17.1 - 20.9] 20.9
96
[5.1 - 14.4] 9.5
25.0
20.0
15.0
10.0
5.0
0.0
MYA

Figure 6



Host taxa (plant family)	
Uroleucon [20]	Asteraceae
Macrosiphoniella [2]	Asteraceae
Aulacorthum [3]	Asteraceae
Macrosiphum [4]	Rosaceae + others
Sitobion [3]	Rosaceae, Poaceae
Metopolophium dirhodum	Rosaceae, Poaceae
Acyrthosiphon pisum	Fabaceae, Rosaceae
Acyrthosiphon chelidonii	Fabaceae, Rosaceae
Hyperomyzus [2]	Asteraceae, Grossulariaceae
Nasonovia ribis-nigri	Grossulariaceae + others
Obtusicauda coweni	Asteraceae
Delphiniobium hanla	Ranunculaceae
Hydronaphis impatiens	Many families
Amphicercidus japonicus	Caprifoliaceae
Megoura [3]	Fabaceae, Rosaceae
Myzus persicae	Many families
Myzus varians	Many families
Dysaphis rumecicola	Many families
Myzus cerasi	Many families
Dysaphis plantaginea	Many families
Brachycaudus [3]	Many families
Brevicoryne brassicae	Brassicaceae
Hayhurstia atriplicis	Amaranthaceae
Lipaphis pseudobrassicae	Brassicaceae
Diuraphis noxia	Poaceae
Hyadaphis tataricae	Caprifoliaceae
Aphthargelia symphoricarpi	Caprifoliaceae
Aphis [2]	Many families
Toxoptera citricida	Rutaceae
Toxoptera aurantii	Rutaceae
Schizaphis graminum	Poaceae + others
Rhopalosiphum padi	Poaceae + others
Pterocomma pilosum	Salicaceae
Cavariella salicicola	Salicaceae
Capitophorus hudsonicus	Asteraceae, Polygonaceae
Prociphilus fraxinifolii	Many families

Figure 7


Host alternation
Alternating
Simple
Both traits present
Uroleucon
U+A clade
Macrosiphini
Aphidinae
Aphidini
Uroleucon [20]
Macrosiphoniella [2]
Aulacorthum [3]
Macrosiphum [4]
Sitobion [3]
Metopolophium dirhodum
Acyrthosiphon pisum
Acyrthosiphon chelidonii
Hyperomyzus [2]
Nasonovia ribis-nigri
Obtusicauda coweni
Delphiniobium hanla
Hydronaphis impatiens
Amphicercidus japonicus
Megoura [3]
Myzus persicae
Myzus varians
Dysaphis rumecicola
Myzus cerasi
Dysaphis plantaginea
Brachycaudus [3]
Brevicoryne brassicae
Hayhurstia atriplicis
Lipaphis pseudobrassicae
Diuraphis noxia
Hyadaphis tataricae
Aphthargelia symphoricarpi
Aphis [2]
Toxoptera citricida
Toxoptera aurantii
Schizaphis graminum
Rhopalosiphum padi
Pterocomma pilosum
Cavariella salicicola
Capitophorus hudsonicus
Prociphilus fraxinifolii

Chapter 3 figure captions

Figure 1. Frequency distribution of sequence lengths
Distributions are given for *U. ambrosiae* contigs (n=98,441), *A. pisum* mRNA sequences (n=36,961), *U. ambrosiae* coding sequences (n=53,760), and the subset of contigs with a BLAST hit to a known protein (n=52,027). *Ua* contigs have an excess of short sequences and *Ua* CDS have a deficit, relative to *A. pisum* mRNAs. *Ua* contigs with BLAST hits, however, approximate the distribution of the *A. pisum* sequence set.

Figure 2. Ortholog hit ratio plots for Ua coding sequences
Ua CDS with matches to *Ap* coding sequences were used to calculate the ortholog hit ratio (OHR = (length of Ua CDS)/(length of Ap CDS)). Plots have OHR on the y-axis and *Ap* CDS length on the x-axis. For each pair of sequences, an OHR of 1.0 is indicative of a full-length *Ua* CDS ortholog. Although a large number of short sequences have low OHR values, a considerable number of short sequences are at or near OHR=1, suggesting caution in excluding short sequences from the *Ua* assembly.

Figure 3. Species identity of top BLAST hits
Pie chart of the species identity of top BLAST hits ($e < 1x10^{-5}$) of *Ua* contigs compared to the NCBI non-redundant protein database. 52,027 *Ua* contigs had at least one BLAST hit.

Figure 4. q-value distribution for fold change estimates of expressed transcripts
The q-value distribution was calculated for expression profiles for all coding sequences with minimum read coverage of 10. Based on the false discovery rate, q-values provide a measure of the expected number of false positives associated with specific significance cutoffs. a) Plot of q-value versus p-value. The q-value (x-axis) increases significantly as p-value cutoffs (y-axis) are relaxed. A cutoff of $p<0.004$, accepted as significant in many studies, entails a 20% (q=0.20) rate of acceptance of false positives. We use a cutoff of $p<.0005$, corresponding to q=0.054. b) Plot of the number of false positives (y-axis) as a function of the number of transcripts accepted ("significant tests", x-axis). The false positive rate increases dramatically at cutoffs yielding more than around 125 significantly differentially expressed sequences. We accept 112 transcripts as significantly differentially expressed.

Figure 5. Correlation of expression profiles across RNA-seq samples
Tree based on hierarchical clustering of Spearman correlations of change in normalized expression values for each transcript (rows) in each treatment (columns). Rows on the x-axis shows each of the 124 differentially expressed genes. a) Tree depicting hierarchical similarity relationships of treatments based on expression profiles across all differentially expressed transcripts. Treatments that group together have similar expression profiles. b) Tree depicting hierarchical similarity relationships of transcripts based on correlation of expression profiles across the six samples. Transcripts that group together have expression profiles that are similarly affected by host plant. c) Clustering of transcripts with similar expression levels at a cutoff of nodes above 20% of the total length of the tree—each color represents a cluster. d) Heat map illustrating relative expression levels

for each transcript (rows) in each sample (columns). Adjacent (right to left) bands of the same color indicate that the given transcript had a similar expression level in response to the given treatment. Overall, expression profiles of the two IF replicates are highly correlated to each other. This is also true for the AT replicates. The TR replicates produced divergent expression profiles. IF = *Iva frutescens,* TR = *Tithonia rotundifolia,* AT = *Ambrosia trifida*.

Figure 6. Representative CDS clusters with correlated expression profiles

Clusters of *Ua* coding sequences with divergent correlated co-expression profiles in response to host plant treatment. 91 unique *Ua* sequences are members of clusters with divergent correlated expression profiles on AT versus IF (x-axis); two such clusters are shown here. a) Cluster including five isogenes involved in purine metabolism. This cluster illustrates a case where expression profiles are consistent within each host plant treatment. b) Example of cluster where expression profiles are consistent within AT and IF treatments, but not TR (shaded middle region). Results for TR were discarded due to consistently divergent profiles across *Ua* CDS among TR replicates. Correlations are based on estimated fold change (y-axis) at cutoffs of fold change >2, $p<0.0005$. Light gray lines represent individual CDS profiles and blue lines indicate the line of best fit, *i.e.* correlated expression profile. AT = *Ambrosia trifida,* TR = *Tithonia rotundifolia,* IF = *Iva frutescens*.

Figure 7. Examples of the three *Ua*-only clade classes

Representative cladograms based on ML analysis of phylome multiple sequence alignments. Each cladogram represents the relationship of *Ua* CDS to their most closely related insect protein from a non-*Ua* species. The number of *Ua* CDS that are monophyletic with respect to a non-*Ua* sister are used to classify each clade into one of three *Ua*-only group sizes: a) One-to-one relationship of *Ua* coding sequence to *A. pisum* protein. b) Few-to-one relationships (between 2-9 monophyletic *Ua* sequences c) Many-to-one relationship (over 10 monophyletic *Ua* sequences). All *Ua* coding sequence IDs are shown in blue. IDs with suffixes (e.g. −1, −2) represent multiple open reading frames on a single transcript. Codes for non-*Ua* genes are as follows: ACYPI = *Acyrthosiphon pisum,* ANOGA = *Anopheles gambiae*, DAPPU = *Daphnia pulex*, DROME = *Drosophila melanogaster,* IXOSC = *Ixodes scapularis*, NASVI = *Nasonia vitripennis*, TRICA = *Tribolium castaneum*.

Figure 8. Linear regression of group size on length for Ua-only phylome groups

Plot of length distribution and linear model best fit lines for Ua-only group size against mean sequence lengths for each group. Best fit lines are given for the total set of Ua-only groups (black), as well as many-to-one (red) and few-to-one (green) groups. According to regression analysis, the correlation coefficient for many-to-one groups is not different from few-to-one groups or the total groups ($p>0.1$). Few-to-one groups have a significantly higher correlation the groups as a whole ($p<0.05$). Note that the y-axis is relatively zoomed to show the distribution of group sizes; proportional axes would result in much flatter lines at these R^2 values. Transcript length is not a significant factor in the frequency of many-to-one Ua-only groups.

Figure 9. Frequency distribution of OHR values among Ua-only phylome groups
Density functions for ortholog hit ratio (OHR = (length of Ua CDS)/(length of Ap CDS)) values for the transcripts assigned to one-to-one (blue), few-to-one (green), and many-to-one *Ua*-only clades. The plots indicate roughly bimodal distributions (trimodal for one-to-one groups) with major peaks around OHR=1. Based on this distribution, Ua sequences with OHR<0.75 were excluded from downstream analysis to focus on complete and nearly complete coding sequences. Density plots were generated using a biweight kernel to build distributions around each data point, and then smoothing over the peaks of the distributions using the default nrd0 method provided by the R density() function. Density plots enable comparison of distributions by standardizing frequency distributions such that the total density is equal to one.

Figure 10. Frequency distribution of *Ua*-only clade sizes
Ua-only clades are monophyletic groups of *Ua* sequences that are sister to a non-*Uroleucon* species in phylome trees. The x-axis plots *Ua*-only clade sizes from 1 to 50, based on the number of *Ua* sequences in each clade. There were a total of 17,826 *Ua*-only clades, populated by 31,112 *Ua* CDS. Individual *Ua* CDS may be placed in more than one *Ua*-only clade. a) The y-axis gives the frequency of each *Ua*-only clade size. In total, 31,112 *Ua* CDS were placed in 17,826 *Ua*-only groups. The major portion (84%) of CDS are in group sizes under 10. b) Same distribution as in a), but with a narrower y-axis range in order to zoom in on the right tail of the distribution. In examining evidence for lineage-specific gene duplication, we focus on clade sizes over 10 (the right tail).

Figure 1

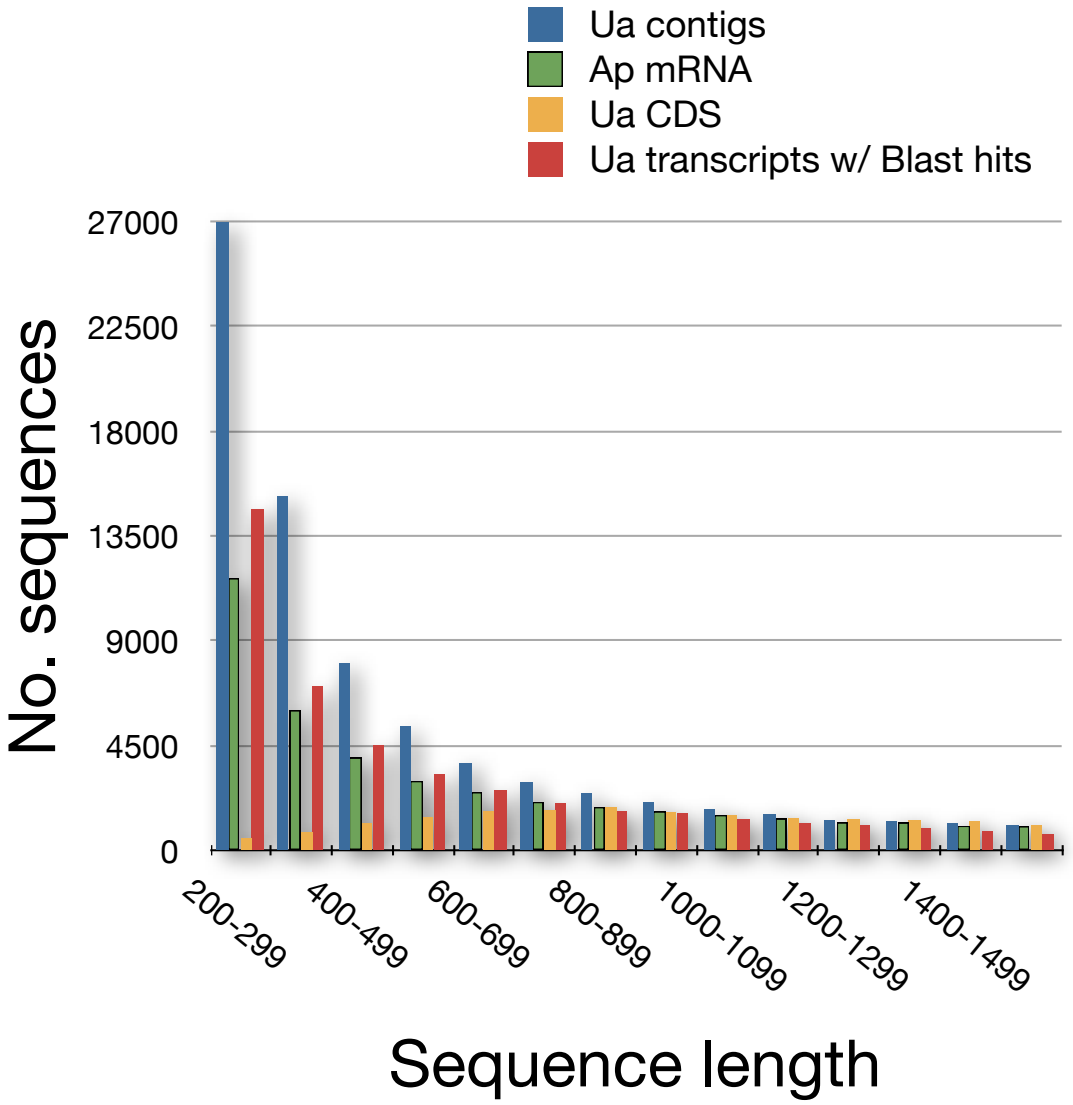

Ua contigs
Ap mRNA
Ua CDS
Ua transcripts w/ Blast hits
27000
22500
18000
13500
9000
4500
0
No. sequences
200-299
400-499
600-699
800-899
1000-1099
1200-1299
1400-1499
Sequence length

Figure 2

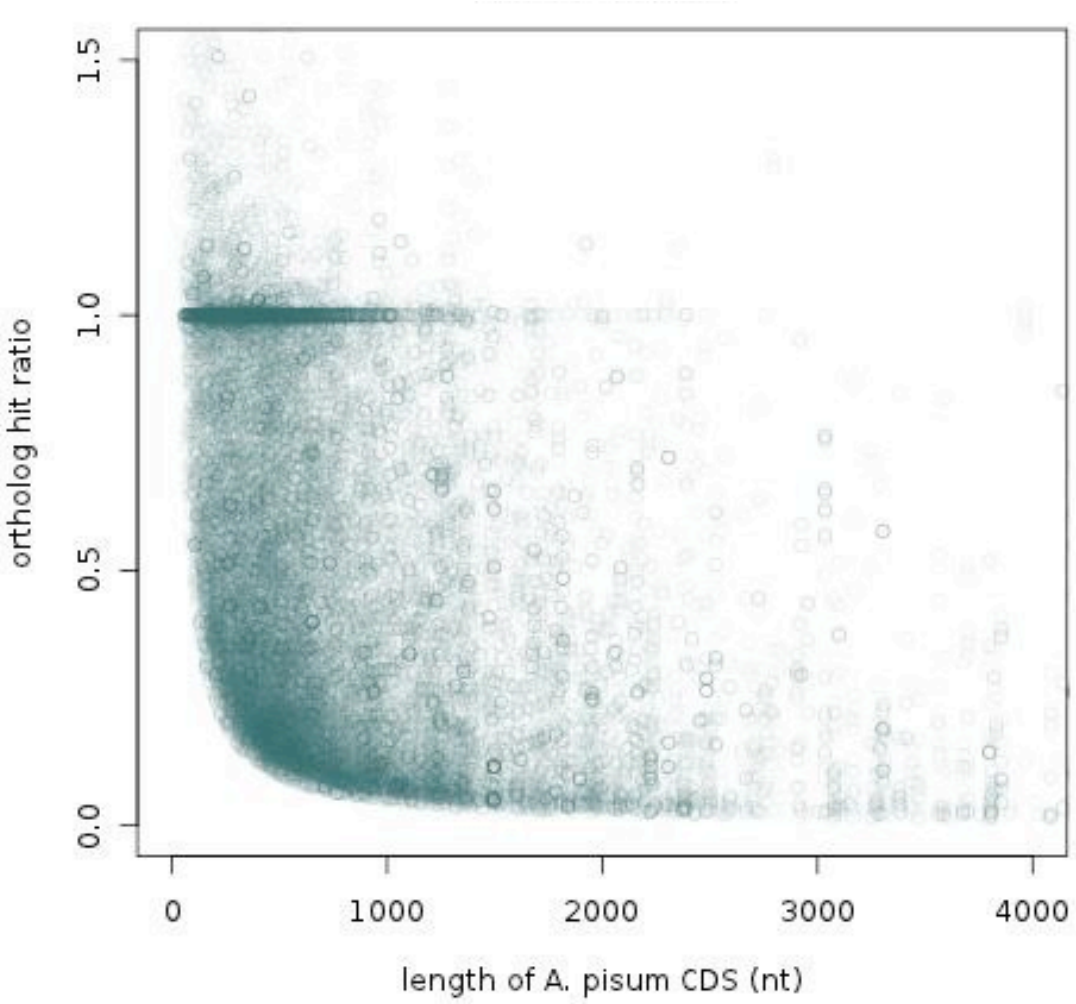

ortholog hit ratio
0.0
0.5
1.0
1.5
0
1000
2000
3000
4000
length of A. pisum CDS (nt)

Figure 3

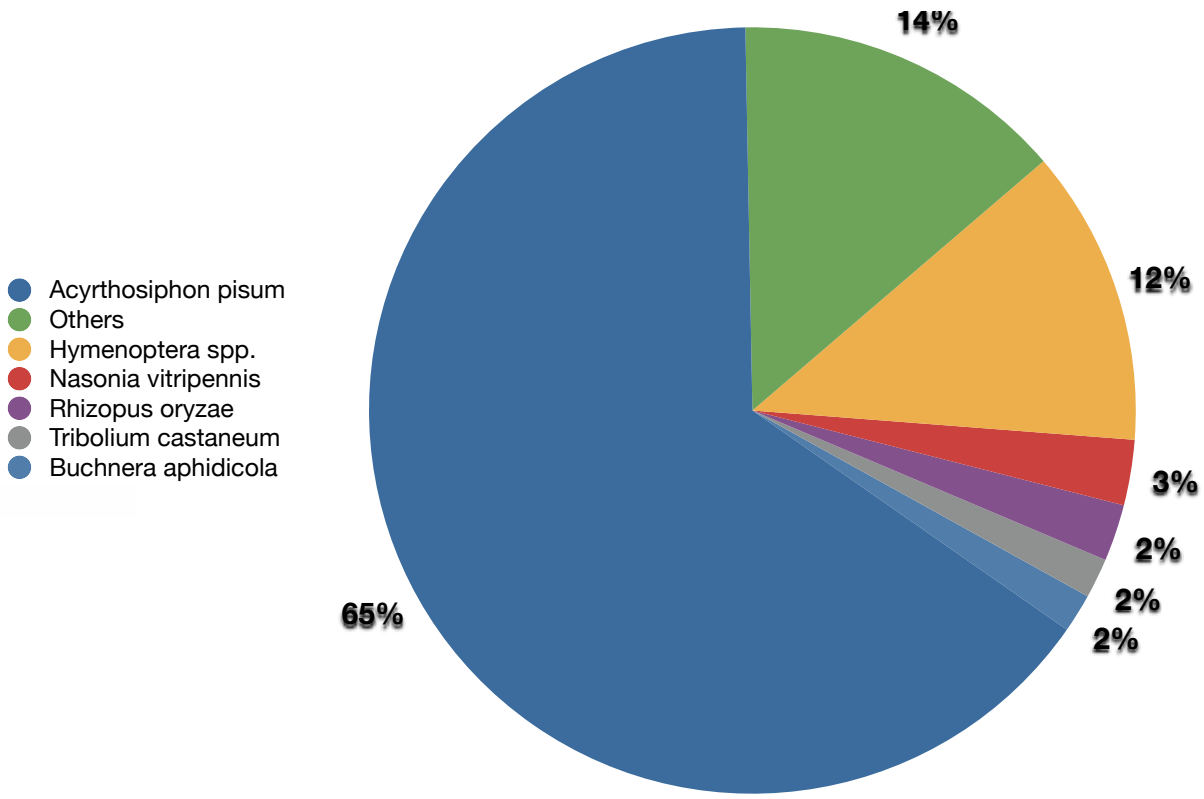

Acyrthosiphon pisum
Others
Hymenoptera spp.
Nasonia vitripennis
Rhizopus oryzae
Tribolium castaneum
Buchnera aphidicola
14%
12%
3%
2%
2%
2%
65%

Figure 4

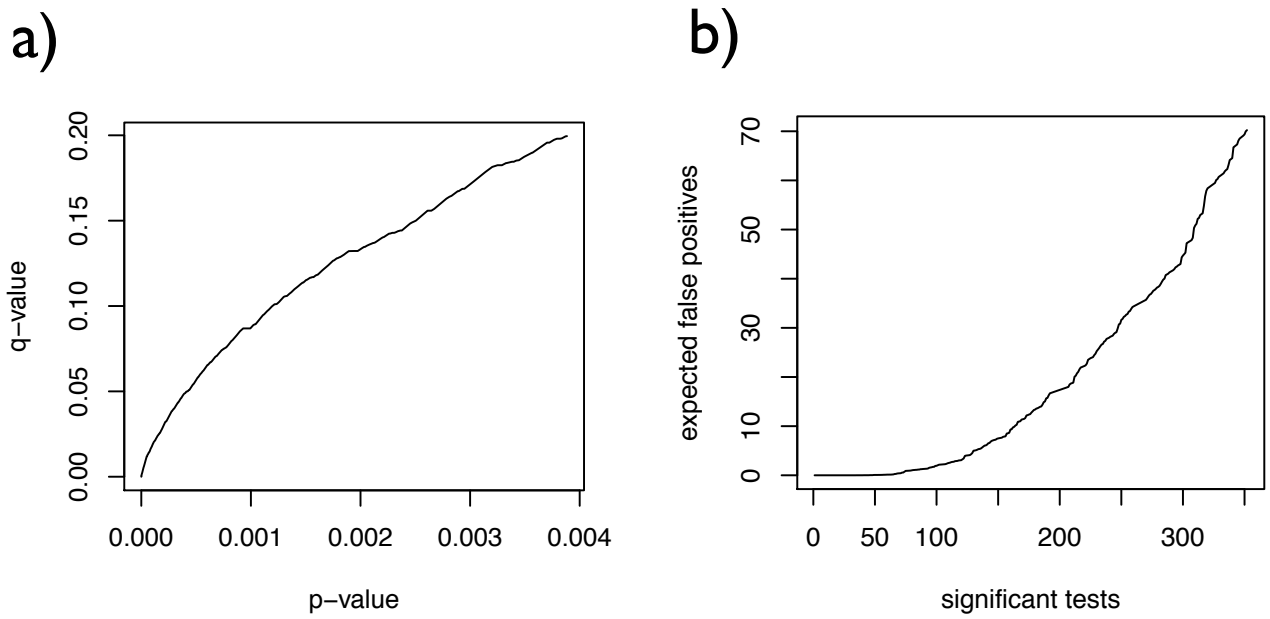

a)
q−value
0.00
0.05
0.10
0.15
0.20
0.000
0.001
0.002
0.003
0.004
p−value
b)
expected false positives
0
10
30
50
70
0
50
100
200
300
significant tests

Figure 5

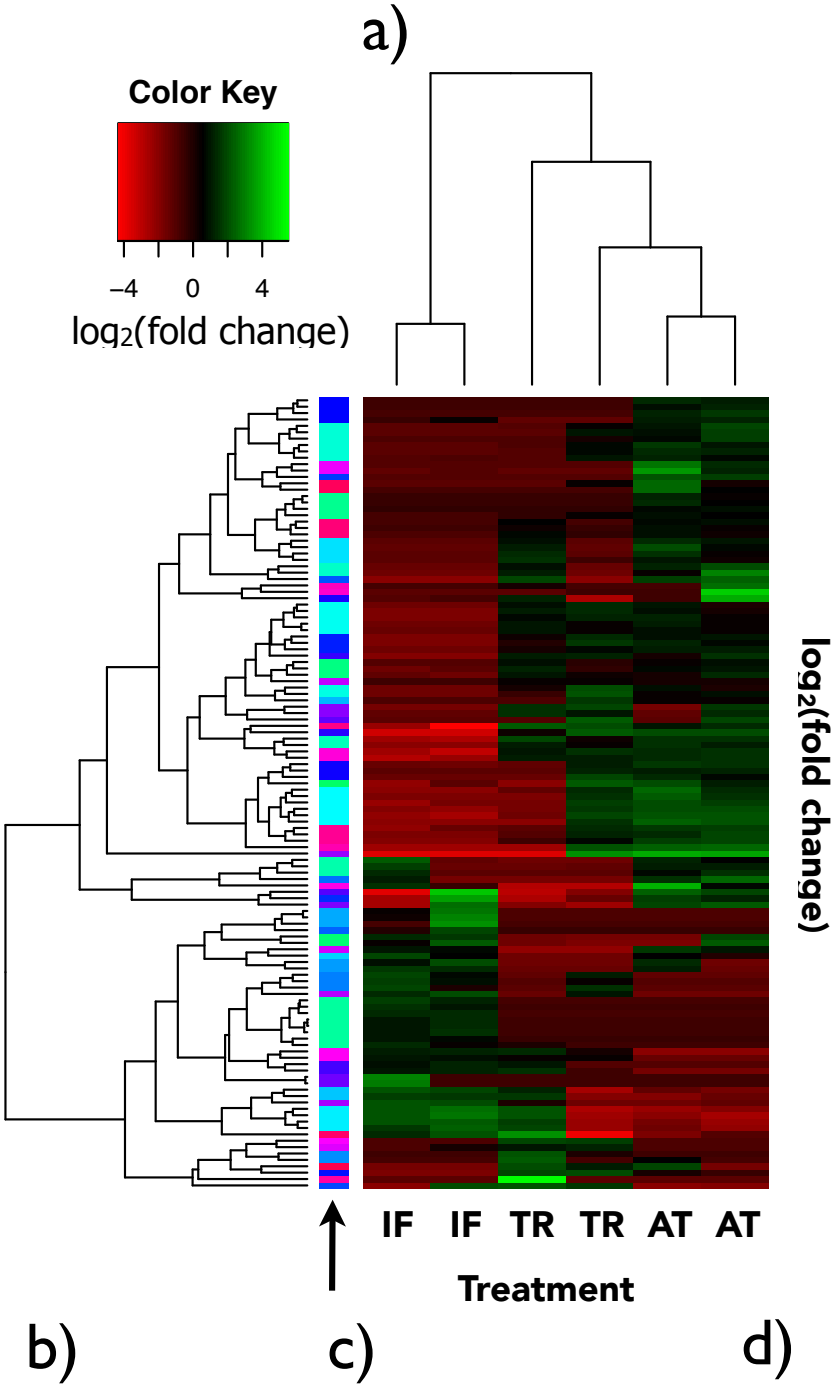

a)
Color Key
−4
0
4
log2(fold change)
log2(fold change)
IF
IF
TR
TR
AT
AT
Treatment
b)
c)
d)

Figure 6

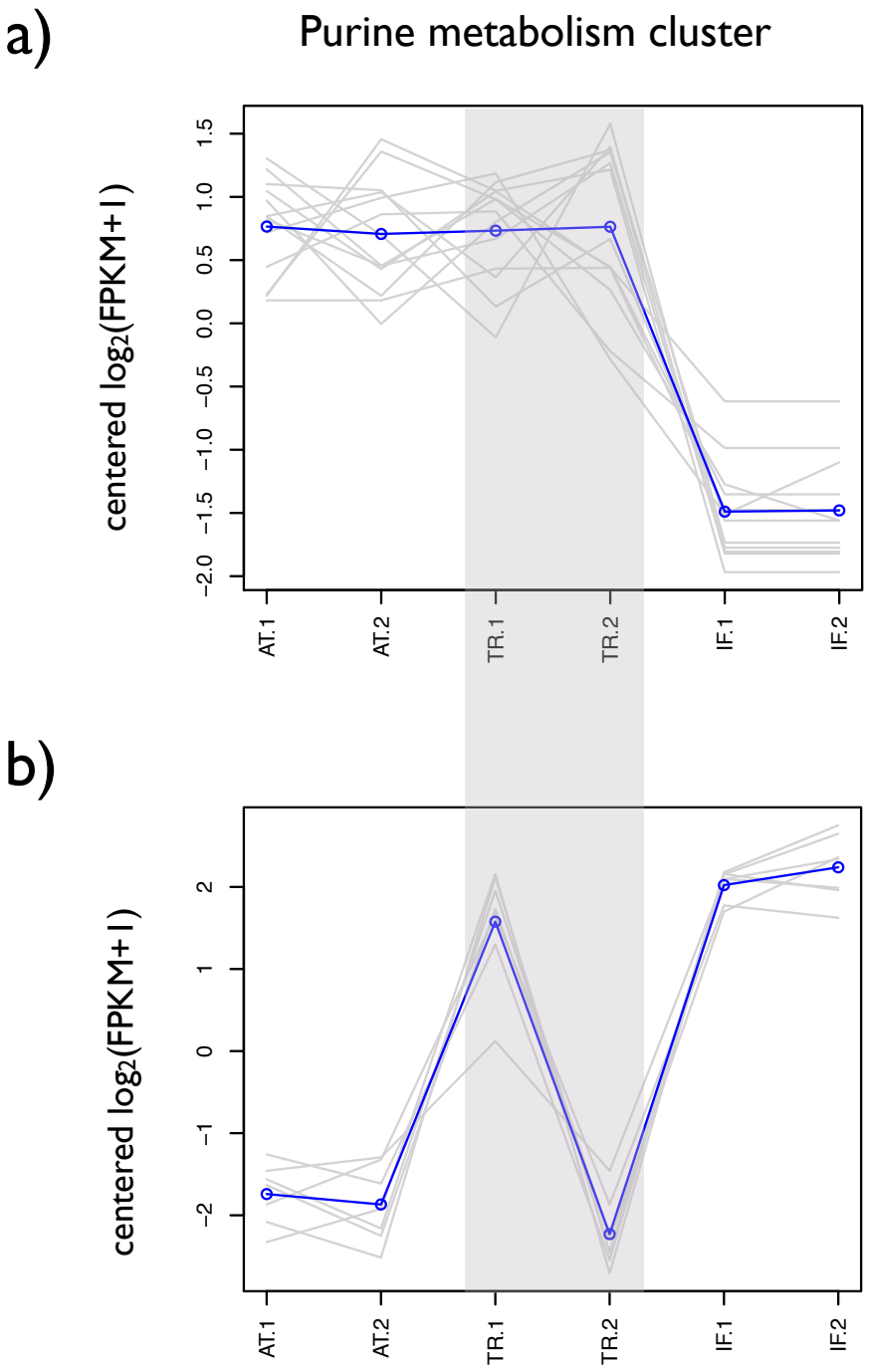

a)
Purine metabolism cluster
centered log2(FPKM+1)
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
AT.1
AT.2
TR.1
TR.2
IF.1
IF.2
b)
centered log2(FPKM+1)
−2
−1
0
1
2
AT.1
AT.2
TR.1
TR.2
IF.1
IF.2

Figure 7

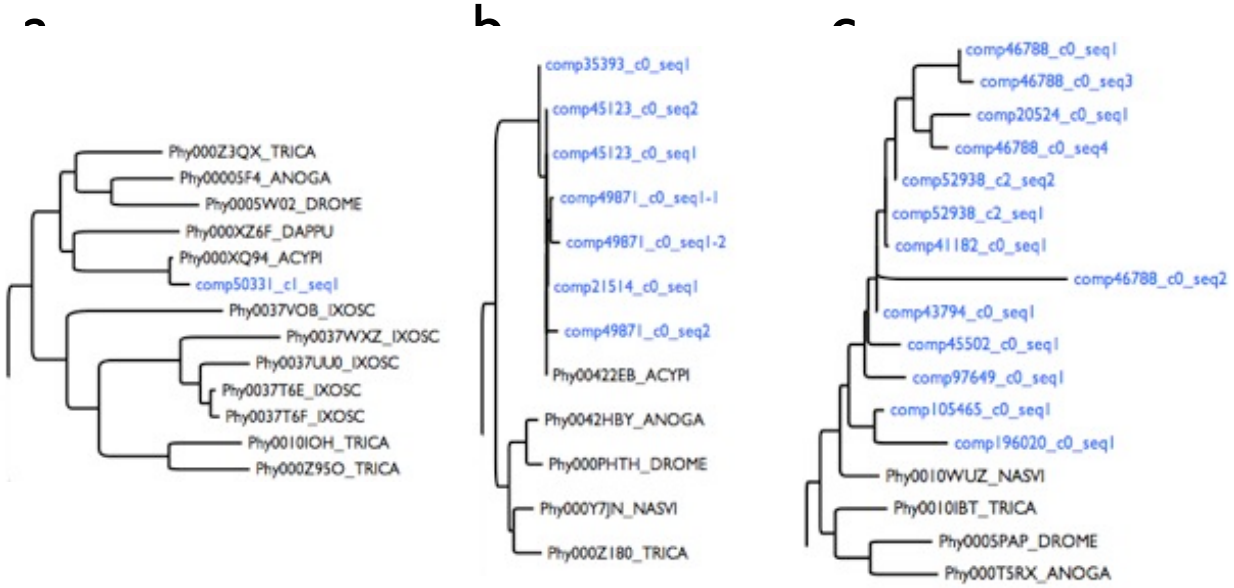

a
Phy000Z3QX_TRICA
Phy00005F4_ANOGA
Phy0005W02_DROME
Phy000XZ6F_DAPPU
Phy000XQ94_ACYPI
comp50331_c1_seq1
Phy0037VOB_IXOSC
Phy0037WXZ_IXOSC
Phy0037UU0_IXOSC
Phy0037T6E_IXOSC
Phy0037T6F_IXOSC
Phy0010IOH_TRICA
Phy000Z95O_TRICA
b
comp35393_c0_seq1
comp45123_c0_seq2
comp45123_c0_seq1
comp49871_c0_seq1-1
comp49871_c0_seq1-2
comp21514_c0_seq1
comp49871_c0_seq2
Phy00422EB_ACYPI
Phy0042HBY_ANOGA
Phy000PHTH_DROME
Phy000Y7JN_NASVI
Phy000Z180_TRICA
c
comp46788_c0_seq1
comp46788_c0_seq3
comp20524_c0_seq1
comp46788_c0_seq4
comp52938_c2_seq2
comp52938_c2_seq1
comp41182_c0_seq1
comp46788_c0_seq2
comp43794_c0_seq1
comp45502_c0_seq1
comp97649_c0_seq1
comp105465_c0_seq1
comp196020_c0_seq1
Phy0010WUZ_NASVI
Phy0010IBT_TRICA
Phy0005PAP_DROME
Phy000T5RX_ANOGA

Figure 8

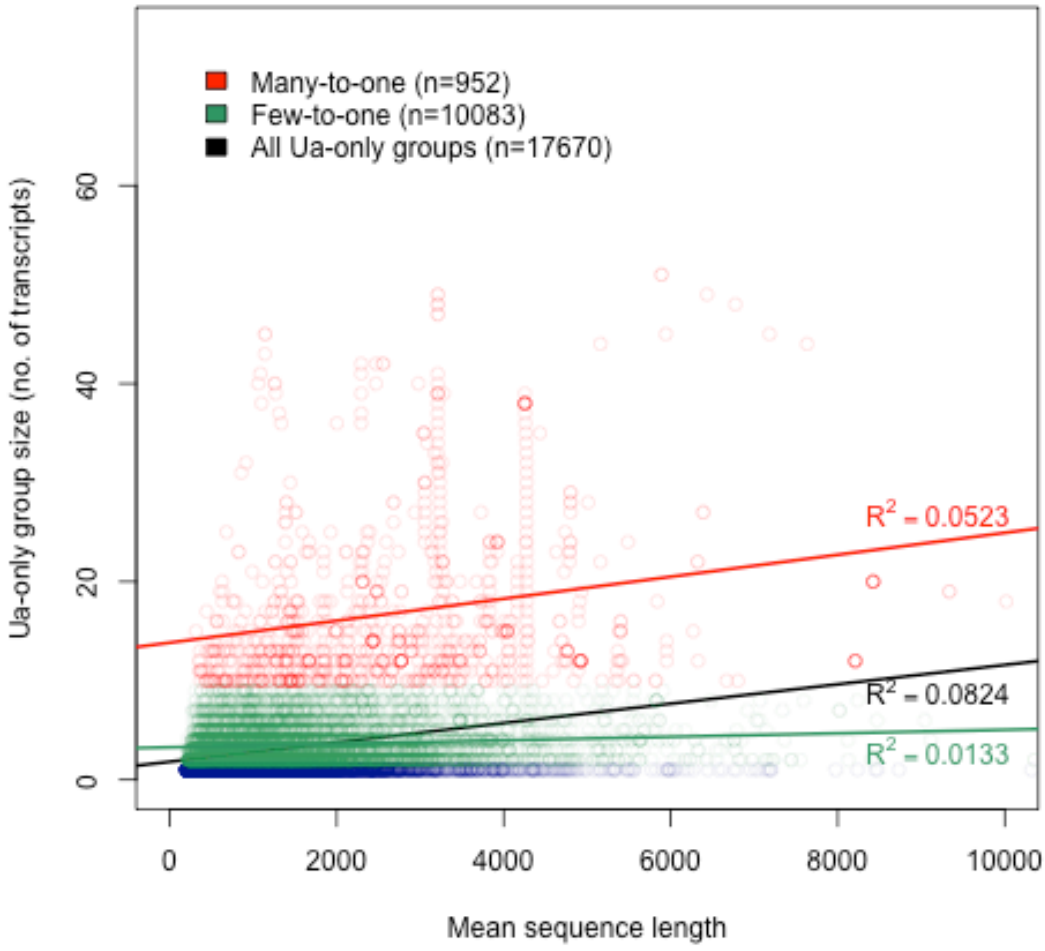

Many-to-one (n=952)
Few-to-one (n=10083)
All Ua-only groups (n=17670)
R² = 0.0523
R² = 0.0824
R² = 0.0133
Ua-only group size (no. of transcripts)
0
20
40
60
0
2000
4000
6000
8000
10000
Mean sequence length

Figure 9

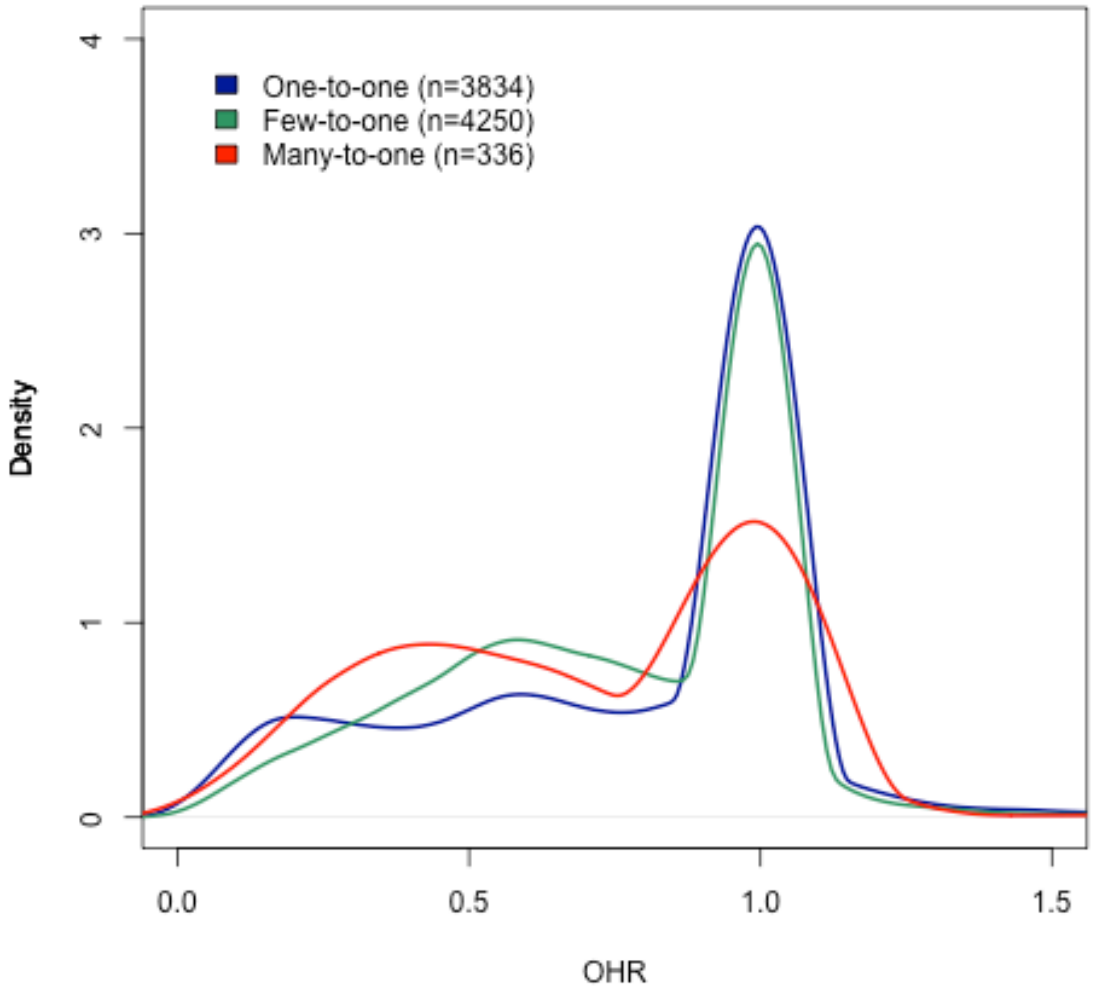
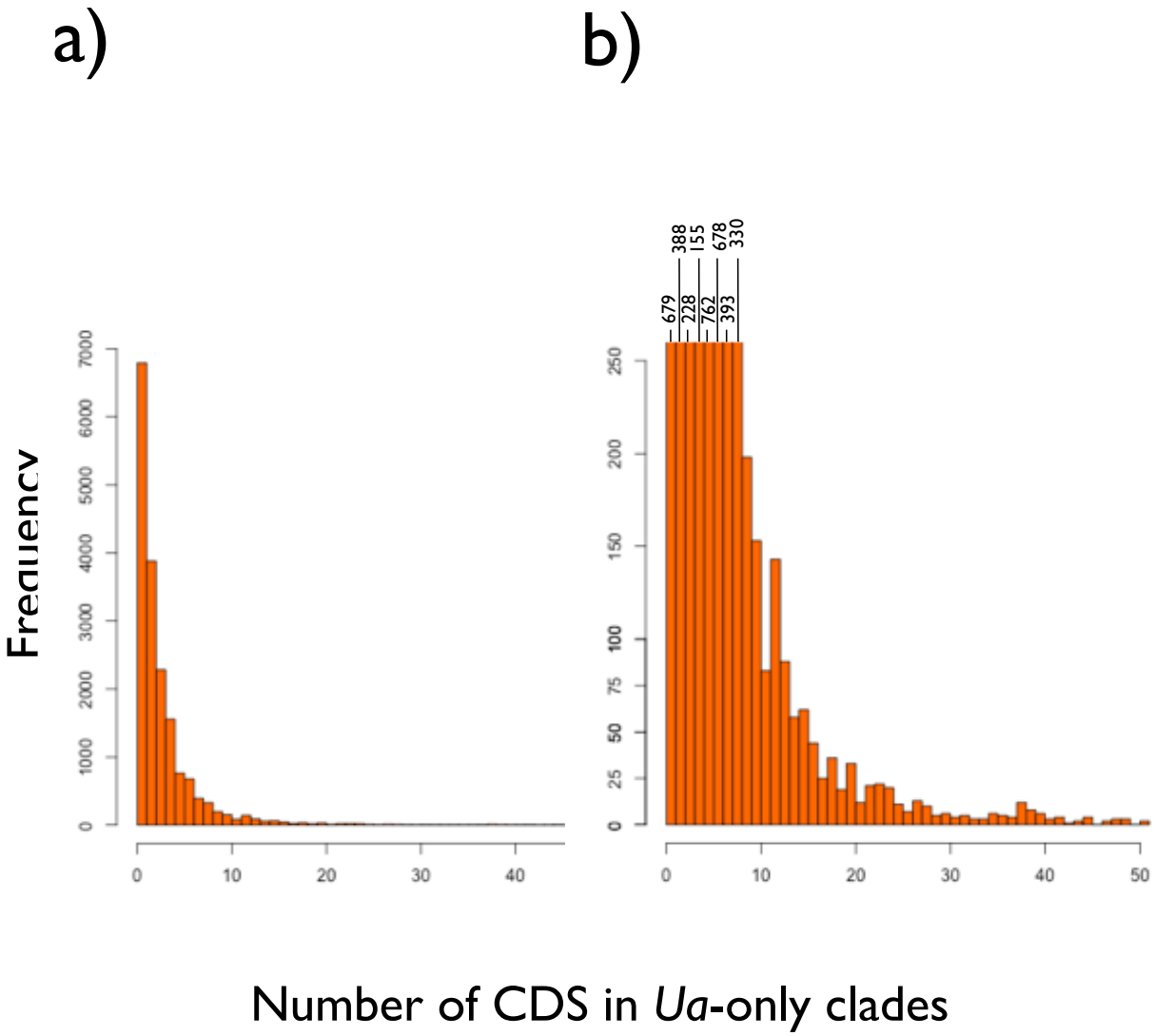

One-to-one (n=3834)
Few-to-one (n=4250)
Many-to-one (n=336)
Density
0
1
2
3
4
0.0
0.5
1.0
1.5
OHR

Figure 10


a)
b)
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
10
20
30
40
679
388
228
155
762
678
393
330
0
25
50
75
100
150
200
250
0
10
20
30
40
50
Number of CDS in *Ua*-only clades

Chapter 4 figure captions

Figure 1. Aphid sampling

Map of populations and source samples used for pooled transcriptome cDNA libraries. From each indicated locality, aphids occupying a single stem or flower were collected into a vial that was placed directly into liquid nitrogen. Following cDNA synthesis, all samples from a given population were pooled. Numbers of samples per population (n) are indicated for each population.

Figure 2. Histogram of per-site sequencing depth for all polymorphic sites

Read depth is determined based on the number of reads aligned at a given nucleotide site. Each bar represents a bin of size 5. The average read depth is 44.9 (n = 371,608 sites).

Figure 3. Frequency distribution of filtered SNPs per isogene

Distribution of 39,245 SNPs that passed quality filtering on 6,240 isogenes. Each bar represents the number of SNPs per isogene.

Figure 4. Distribution of pair-wise allele frequency differences

Pair-wise allele frequency differences, $D = | P_{pop1} - P_{pop2}|$, where P is the estimated allele frequency at a given locus. D is calculated for all 39,245 SNPs. Distributions for each pair-wise comparison are given with kernel density plots, which enable comparison of distributions by standardizing frequency distributions such that the total density is equal to one. Specialist-generalist comparisons are shown in "warm" colors and same-breadth comparisons are shown in blue and green. The density plot depicts a bimodal distribution for D, with the majority of divergence associated with specialist-generalist and not same-breadth comparisons. D values are on the x-axis, and density values are on the y-axis. Density plots were generated using a biweight kernel to build distributions around each data point, and then smoothing over the peaks of the distributions using the default nrd0 method provided by the R density() function.

Figure 5. Frequency distribution of pair-wise F_{ST} values

Pairwise F_{ST} values for 300nt windows. F_{ST} was calculated for all windows with minimum coverage of at least ten sequence reads (n=6242). F_{ST} values are on the x-axis and the calculated density of transcripts at each F_{ST} value is on the y-axis. F_{ST} -density functions are plotted for specialist vs. generalist comparisons ("warm" colors), as well as for same-breadth comparisons (specialist vs. specialist and generalist vs. generalist) in "cool" colors (see legend). Frequencies are illustrated using kernel density plots, which enable comparison of distributions by standardizing frequency distributions such that the total density is equal to one. Density plots were generated using a biweight kernel to build distributions around each data point, and then smoothing over the peaks of the distributions using the default nrd0 method provided by the R density() function.

Figure 6. q-value distribution of K_a/K_s p-values

The q-value distribution was calculated based on p-values testing the null hypothesis $K_a = K_s$ using Fisher's exact test (n = 4737). Based on the false discovery rate, q-values provide a measure of the expected number of false positives associated with specific

significance cutoffs. a) Plot of q-value versus p-value. The q-value (x-axis) increases significantly as p-value cutoffs (y-axis) are relaxed. We use a cutoff of $q < 0.006$, corresponding to $p < 0.1$, corresponding to $q=0.054$. b) Plot of the number of false positives (y-axis) as a function of the number of transcripts accepted ("significant tests", x-axis). Notably, q-values are quite low for p-values typically regarded as insignificant, *e.g* $p=0.15$ corresponds to $q=0.008$.

Figure 7. K_a/K_s values versus coding sequence length

K_a/K_s ratios are given on the y-axis for 2043 isogenes. CDS length is on the x-axis. Darker areas of the graph represent greater concentrations of Ka/Ks values at that CDS length. K_a/K_s and length are poorly correlated ($R^2 = 0.002$).

Figure 8. Frequency distribution of pair-wise K_a/K_s values

Pairwise K_a/K_s values (x-axis) for significant SNPs (n=2043). The calculated density, or frequency, K_a/K_s values is on the y-axis. Specialist-generalist comparisons are shown in "warm" colors and same-breadth comparisons are shown in blue and green. Frequencies are illustrated using kernel density plots, which enable comparison of distributions by standardizing frequency distributions such that the total density is equal to one. Density plots were generated using a biweight kernel to build distributions around each data point, and then smoothing over the peaks of the distributions using the default nrd0 method provided by the R density() function.

Figure 9. Differentiation versus geographic distance

Geographic distance between samples in the study plotted against a) mean allele frequency difference (D), and b) mean F_{ST}. Each point in both plots concerns a pair-wise comparison between two samples.

Figure 10. Comparison of outlier results for D, F_{ST}, and K_a/K_s

The four loci at the intersection of these groups represent candidate genes identified in this study, as described in the text.

Figure 1

Figure 2

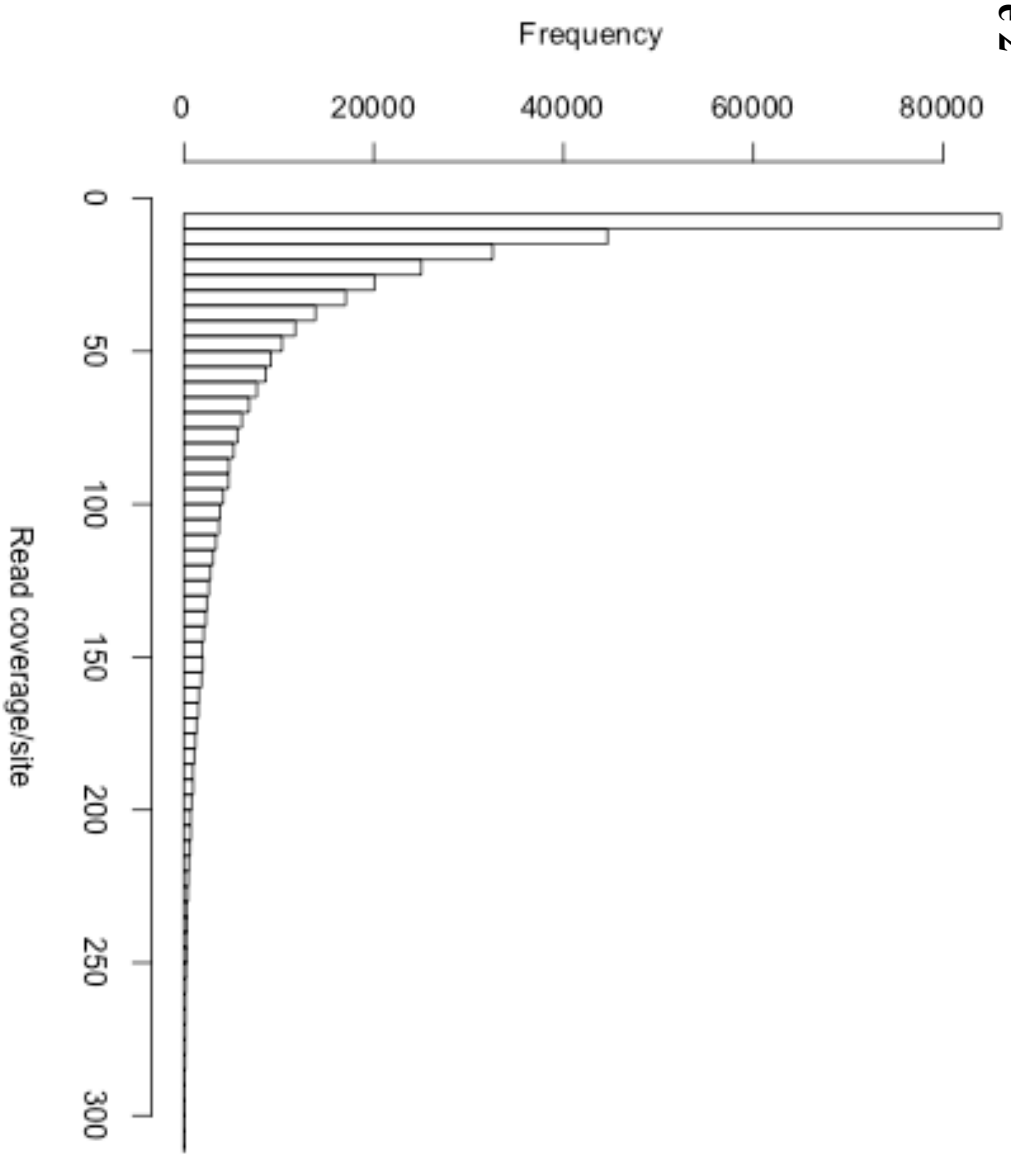

Frequency
0
20000
40000
60000
80000
Read coverage/site
0
50
100
150
200
250
300

Figure 3

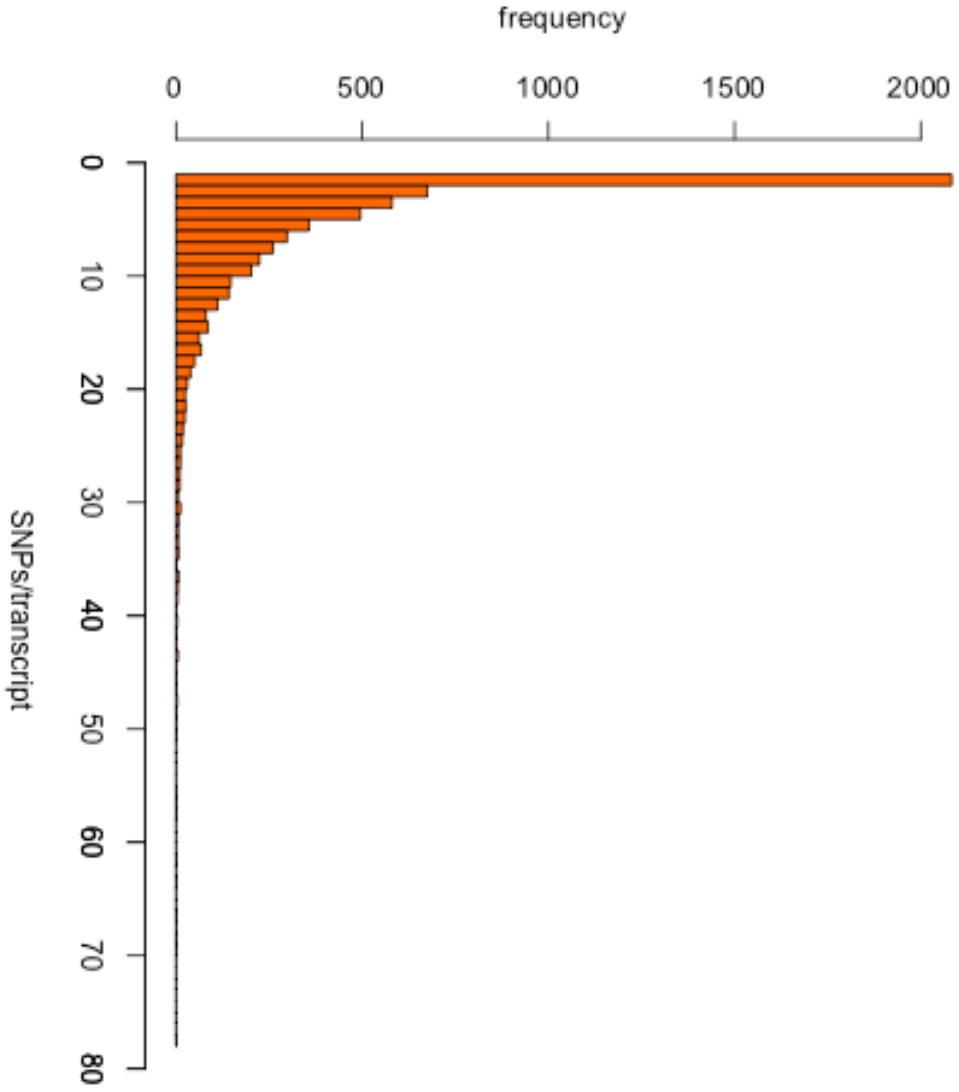

frequency
0
500
1000
1500
2000
SNPs/transcript
0
10
20
30
40
50
60
70
80

Figure 4

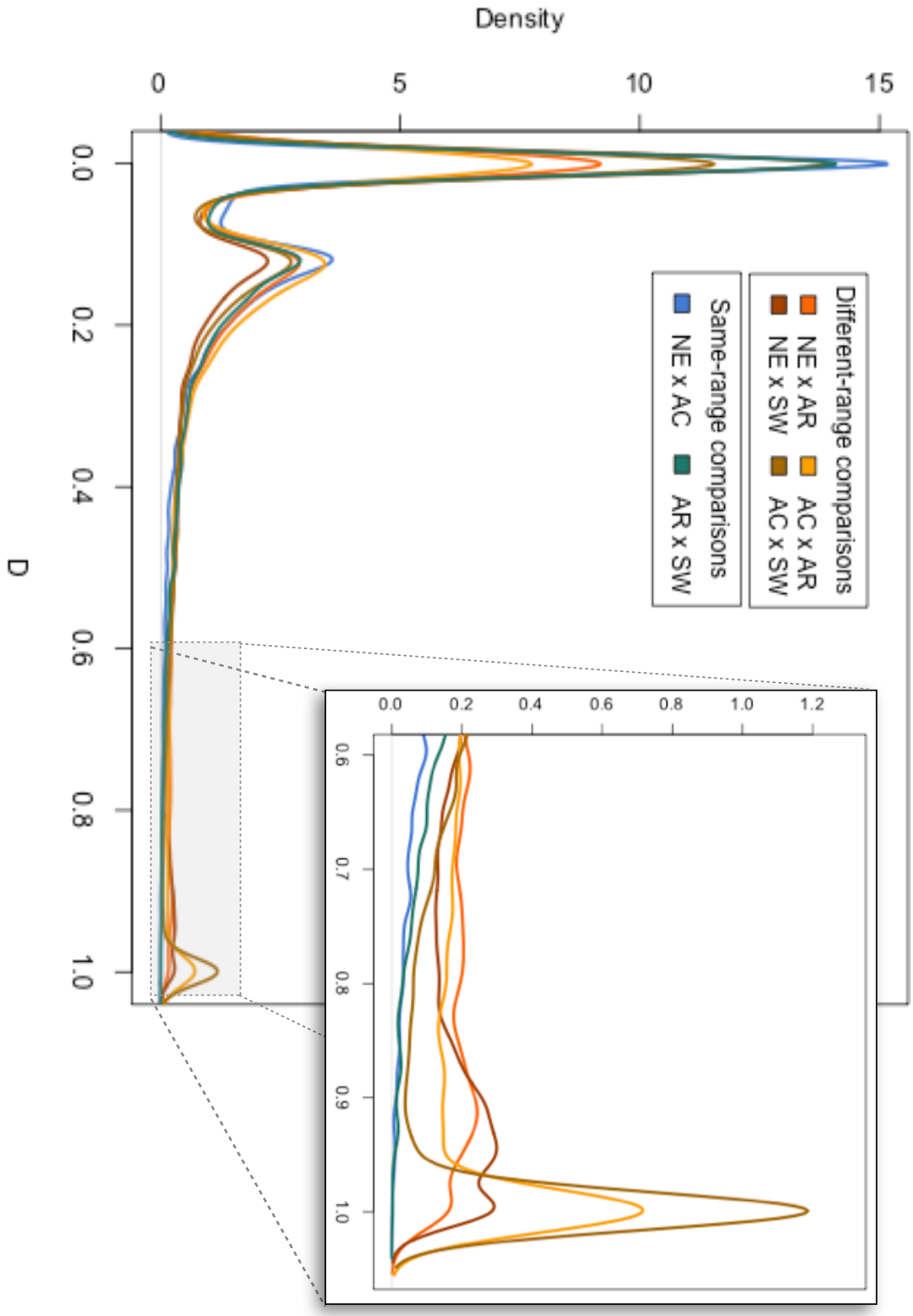

Density
0
5
10
15
D
0.0
0.2
0.4
0.6
0.8
1.0
Different-range comparisons
NE x AR
AC x AR
NE x SW
AC x SW
Same-range comparisons
NE x AC
AR x SW
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.6
0.7
0.8
0.9
1.0

Figure 5

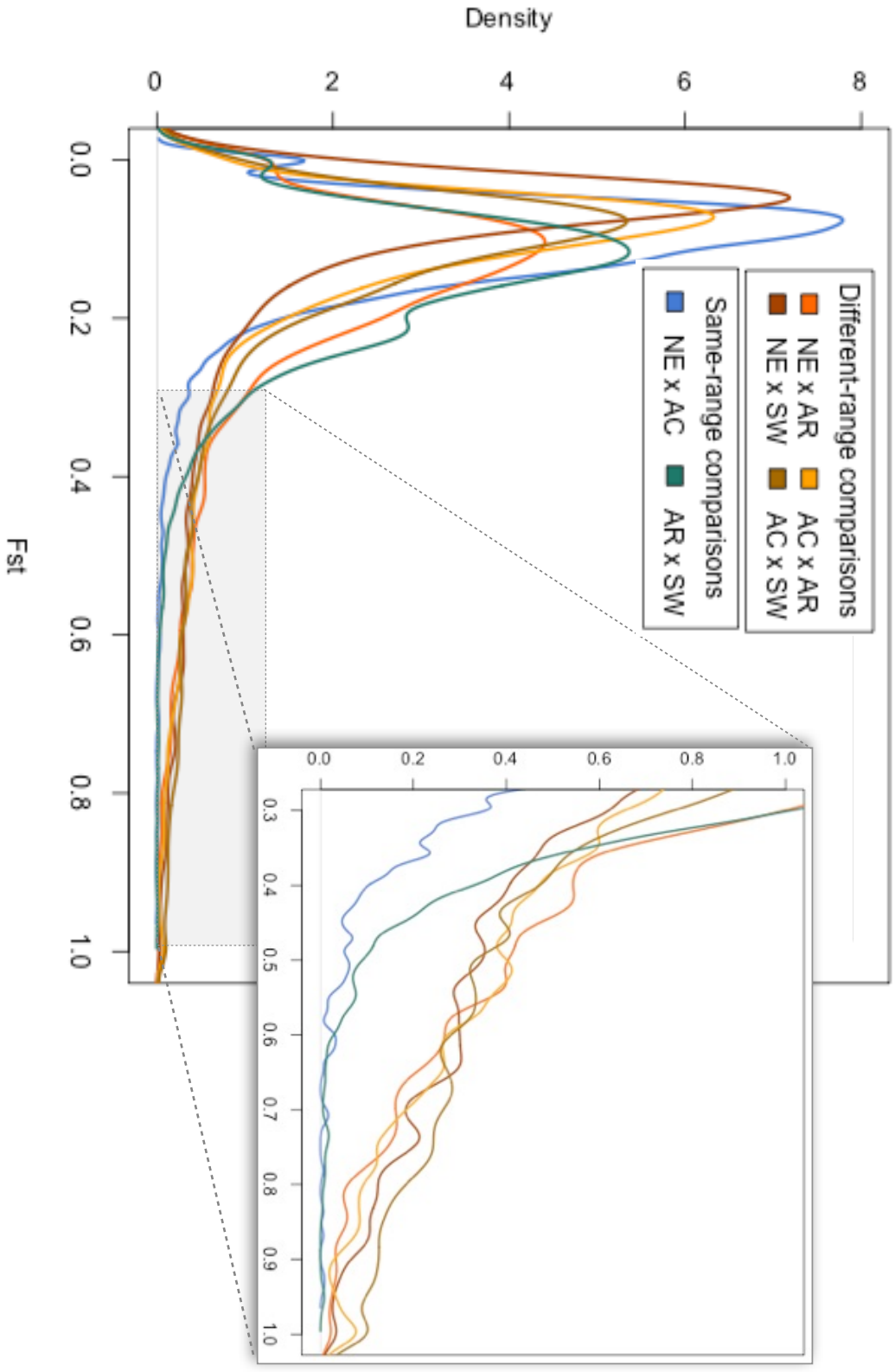

Density
0
2
4
6
8
Fst
0.0
0.2
0.4
0.6
0.8
1.0
Different-range comparisons
NE x AR
AC x AR
NE x SW
AC x SW
Same-range comparisons
NE x AC
AR x SW
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 6

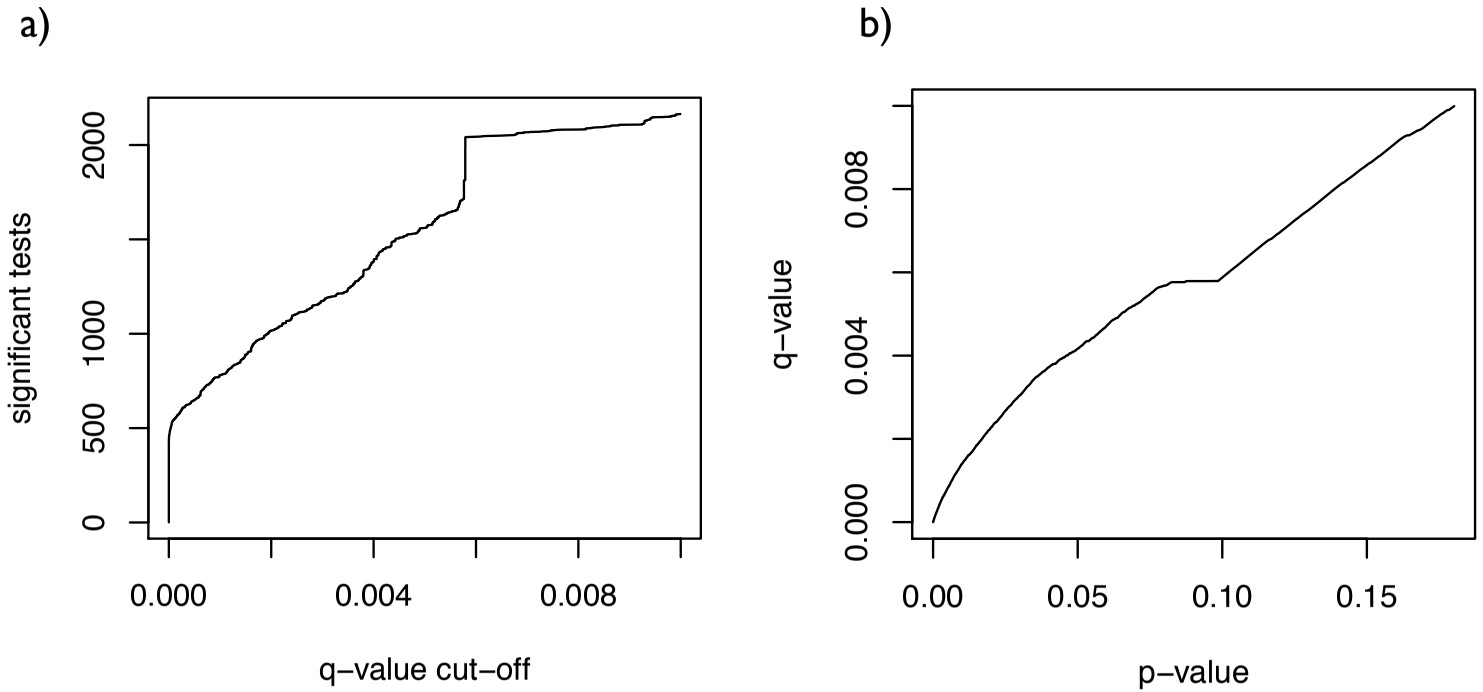

a)
significant tests
0
500
1000
2000
0.000
0.004
0.008
q−value cut−off
b)
q−value
0.000
0.004
0.008
0.00
0.05
0.10
0.15
p−value

Figure 7

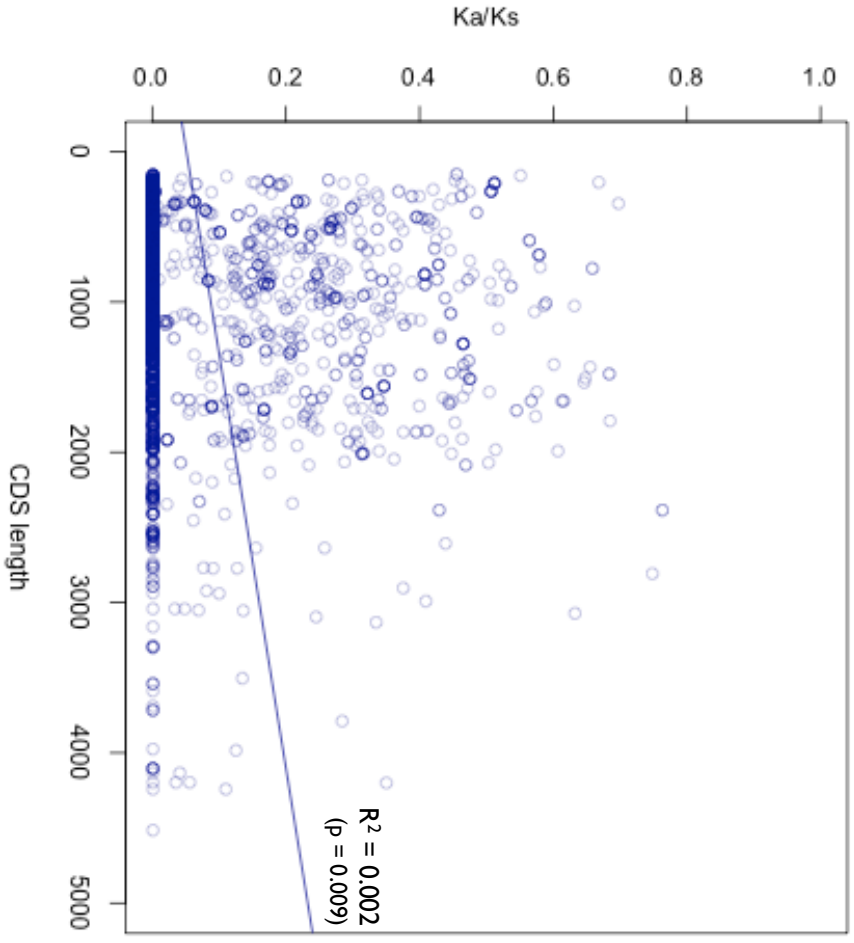

Ka/Ks
0.0
0.2
0.4
0.6
0.8
1.0
CDS length
0
1000
2000
3000
4000
5000
R² = 0.002
(p = 0.009)

Figure 8

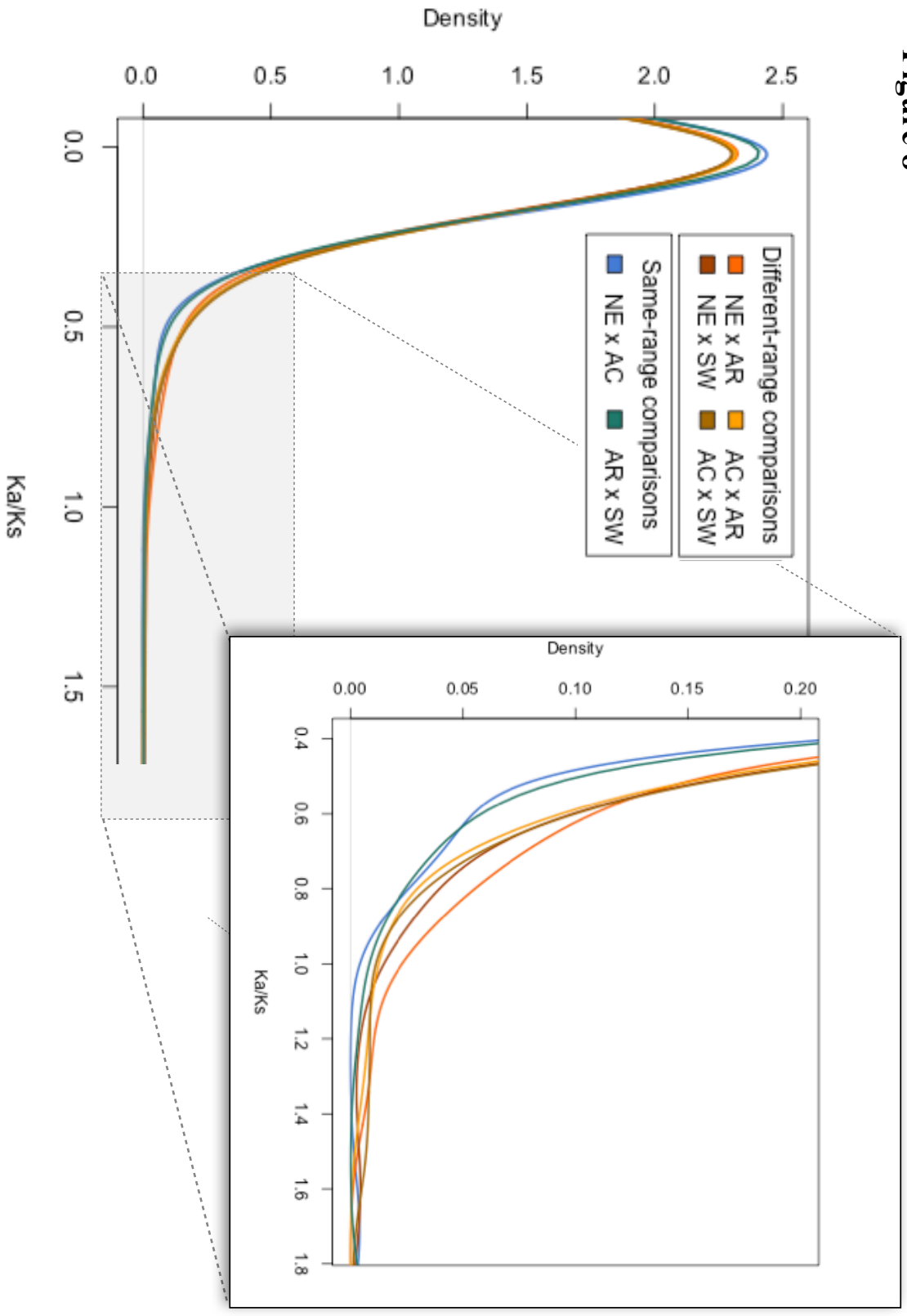

Density
0.0
0.5
1.0
1.5
2.0
2.5
Ka/Ks
0.0
0.5
1.0
1.5
Different-range comparisons
NE x AR
AC x AR
NE x SW
AC x SW
Same-range comparisons
NE x AC
AR x SW
Density
0.00
0.05
0.10
0.15
0.20
Ka/Ks
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8

Figure 9

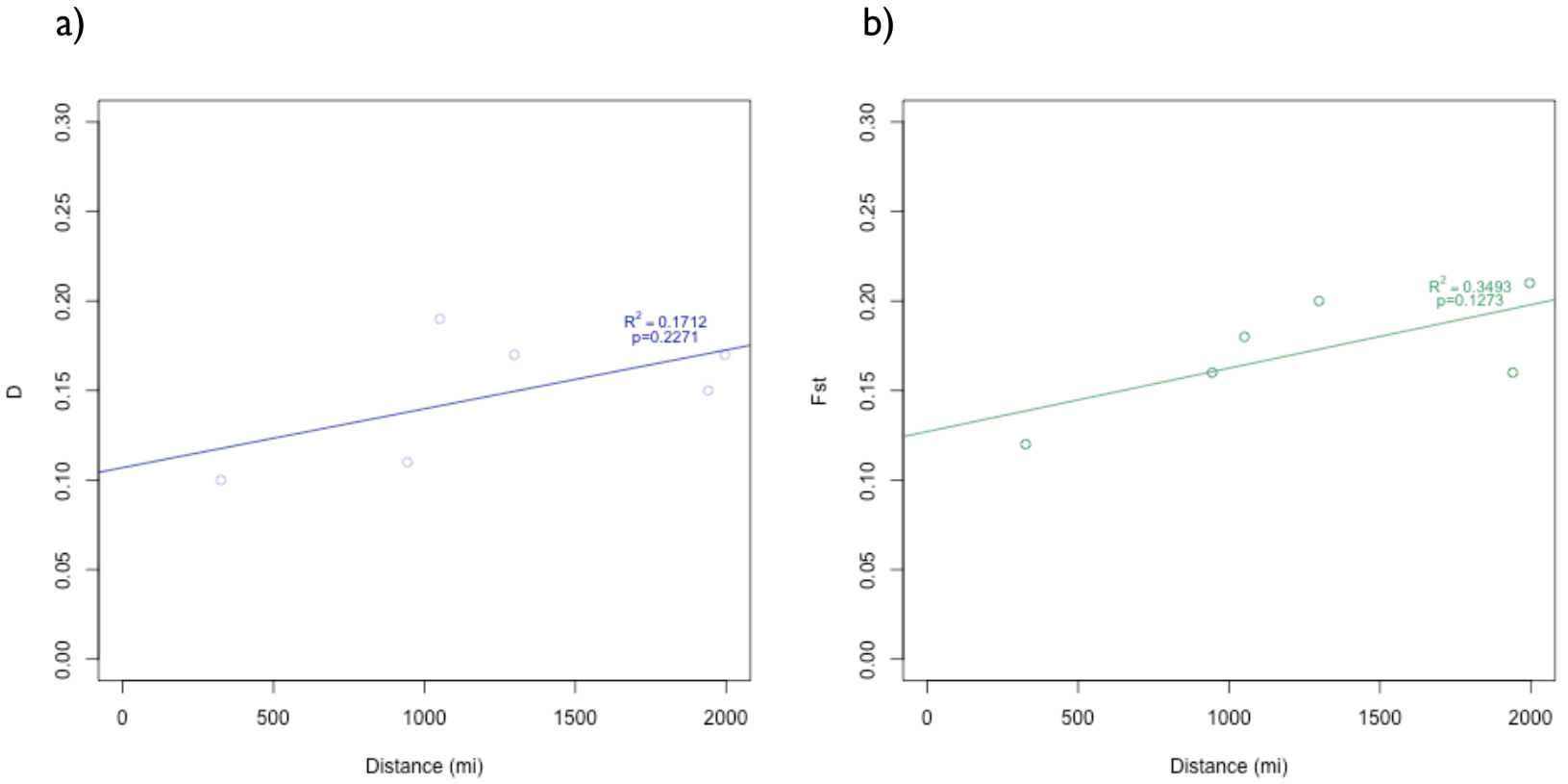

a)
b)
D
Fst
R^2 = 0.1712
p=0.2271
R^2 = 0.3493
p=0.1273
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
500
1000
1500
2000
Distance (mi)
Distance (mi)

Figure 10