HOST PLANT ASSOCIATION AND SPATIAL AUTOCORRELATION AS DRIVERS OF GENETIC DIFFERENTIATION AMONG POPULATIONS OF A REGIONALLY HOST-SPECIFIC INSECT HERBIVORE

By

Amanda L. Driscoe

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
With a Major in Population and Conservation Biology
May 2018

Committee Members:

James R. Ott, Chair

Chris C. Nice

Noland H. Martin

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplicate Permission

ACKNOWLEDGMENTS

The research reported herein was improved by members of both the Population and Conservations Biology Program and the Ecology, Evolution, and Behavior group at Texas State University. This research was funded by Howard McCarley Student Research Awards from the Southwestern Association of Naturalists to Amanda L. Driscoe (ALD), and Robert W. Busbee (RWB); Freeman Center Fellows Awards, and Thesis Research Fellowships from Texas State University to ALD and RWB; a Theodore Roosevelt Memorial Grant from the American Museum of Natural History to RWB and Faculty Research Enhancement Grants in 2013 and 2015 from Texas State University to J. R. Ott. I would like to thank the following contributors for help in obtaining samples: Dr. Gavin Naylor (Charleston, NC), Dr. Evan Brasswell (McAllen, TX), Dr. Carmen Hall (Sapelo Island, GA), Dr. Glen Hood and Linyi Zhang (fall 2016 Louisiana, Mississippi, Alabama, Florida sites).

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation – Description

ANOVA – Analysis of Variance

BP - base pairs

BWA – Burrows-Wheeler Aligner

CI - Cytoplasmic incompatibility

COI – Cytochrome Oxidase Subunit I Gene

CD-HIT – Cluster Database at High Identity with Tolerance

dbMEM – Distance-based Moran's Eigenvector Mapping

dbRDA – Distance-Based Redundancy Analysis

DIC – Deviance Information Criterion

DNA – Deoxyribonucleic acid

ESS – Effective Sample Size

GBS – Genotyping By Sequence

G_{ST} – Nei's 1973 Measure of Genetic Distance

IBD – Isolation by Distance

IBE – Isolation by Environment

ITS2 – Internal Transcribed Spacer Gene

k – Model Parameter, Number of Genomic Clusters

MA – Million Years

MAF - Minor Allele Frequency

MCMC – Markov Chain Monte Carlo

PCA – Principal Components Analysis

PCNM – Principal Coordinates of Neighbor Matrices

PCoA – Principal Coordinate Analysis

PCR – Polymerase Chain Reaction

Qb – Quercus brandegeei

Qf – Quercus fusiformis

Qg – Quercus geminata

Qm – Quercus minima

Qo – Quercus oleoides

Qs – Quercus sagraeana

Qv – Quercus virginiana

RDA – Redundancy Analysis

SNV – Single Nucleotide Variant

SSR – Single Sequence Repeat

q – Estimate Proportion of Ancestry

US – United States

ABSTRACT

Disentangling the processes responsible for structuring patterns of biodiversity at all spatial scales challenges biologists as such patterns represent evolutionary and ecological processes coupled with spatial autocorrelation among sample units. The phytophagous insect, *Belonocnema treatae* (Hymenoptera: Cynipidae) exhibits regional specialization on three species of live oaks throughout its geographic range across the southern USA. Here I ask whether populations of *B. treatae* affiliated with each host plant species exhibit genetic differentiation that parallels host plant phylogeography while controlling for spatial autocorrelation among sampling locations. I used genotyping-by-sequencing of 1,219 *B. treatae* collected from 58 sites distributed across the geographic ranges of the three host plants to identify 6,987 common single nucleotide variants. Population genomic structure was then investigated using a hierarchical Bayesian model to assign individuals to genetic clusters and estimate admixture proportions. To control for spatial autocorrelation when investigating the role of host plant affiliation in determining patterns of among-population genetic differentiation, Distance-based Moran's eigenvector mapping was used to construct regression variables summarizing spatial structure inherent in the sampling design. Redundancy analysis (RDA) incorporating these spatial variables was then used to simultaneously examine the roles of host plant affiliation and spatial autocorrelation in determining patterns of among-population genetic differentiation. Patterns of genomic variation indicate a distinct geographic division east and west of Mississippi, coupled with discrete host

associated lineages in the eastern portion of the species' range and clinal host-associated lineages in the west. RDA confirmed host association as a significant predictor of genomic variation, but longitudinal spatial autocorrelation explained a larger proportion of *B. treatae's* genomic variation. These results suggest *B. treatae* and the host plants share a common evolutionary history that links their patterns of genomic differentiation.

I. INTRODUCTION

Understanding how biodiversity arises is a fundamental question linking ecology, evolution, and genetics (Darwin 1859, Mayr 1942, Bush 1969, Coyne & Orr 2004). Identifying the processes inhibiting gene flow between diverging lineages and the sources of reproductive isolation are critical to uncovering the origins of diversification (Rundle & Nosil 2005). Ecology and geography are often thought as key drivers behind diversification, but are undoubtedly intertwined (Darwin 1859). Classically, geography has received much attention as it provides the spatial foundation for abiotic variation (climate, elevation, soil type, etc.) that can differentially affect ecology across a geographical range. Organisms also have different dispersal capabilities that can vary within and across their biogeographic range; creating reproductive isolation by distance (IBD). Individuals that become geographically isolated from other populations may then experience adaptation and divergent natural selection (Schluter 2001, 2009). This idea has led to the theory of ecological speciation: proposing gene flow is reduced among populations occupying separate specialized niches due to divergent natural selection (Mayr 1942, Bush 1969, Coyne & Orr 2004). Ultimately, continued divergent natural selection can result in the selection of adaptively diverging traits that are genetically correlated to morphological, physiological and (or) behavioral traits linked to reproductive isolation (Schluter 1996, 2009). To demonstrate ongoing ecological speciation there are three required conditions: (1) identification of an ecological source of divergent natural selection, (2) a form of reproductive isolation, and (3) genetic analysis linking the two (Rundle & Nosil 2005). The latter can be used to test for patterns of isolation by environment (IBE) by measuring the extent of genomic differentiation within

and among populations occupying alternate environments. Significant differentiation coupled with patterns of genomic variation that consistently follow environmental partitions (i.e. host association) would give support to ecologically diverging lineages (Mandeville et al. 2015).

Herbivorous insects represent a diverse range of taxa able to overcome of the challenges of host plant defenses to exploit host resources (Mitter et al. 1988, Jaenike 1990). Representing a fourth of macroscopic species, herbivorous insects are among the most ecologically specialized organisms with their extensive diversity stemming from adaptive radiation (Bernays & Chapman 1994, Winkler et al. 2018). Adaptation to new host plant species can promote ecological divergence within herbivores (Funk et al. 2002) which can manifest in the form of differences in habitat preference, temporal isolation, and sexual selection among insect populations occupying alternative host plants (Rundle & Nosil 2005, Servedio 2016). Furthermore, this intimate linkage of biology restricts herbivores biogeography to the geographic range of their host species (Wiklund 1974, Courtney & Forsberg 1988). Herbivorous insects can demonstrate regional host species specialization with the preference for a certain species when multiple suitable species are present (Jaenike 1990). At a finer spatial scale, a herbivores distribution may be patchy due to individual host plant quality related to host genotype and insect population density (Kuussaari et al. 2000, Underwood & Rausher 2000). Regional host species preferences can arise allopatrically if the geographic range of herbivore populations mirrors geographic range expansion/contraction of the host species (Hunter & Price 1992, Underwood & Rausher 2000). Once separated, these herbivore populations may experience differing abiotic and biotic environments that can alter the trajectory of the

arms race between herbivore and host, potentially contributing to reproductive isolation among these adaptively diverging populations.

Upon secondary contact, the extent of historical reproductive isolation can be assessed by examining patterns of genome-wide differentiation (Gompert el al. 2014, Mandeville et al. 2015, Parchman et al. 2016). Genome-wide variation can illustrate patterns of species boundaries as well as population level sub-structuring within species due to adaptive divergence among populations or simply be due to spatial and dispersal limitations (Wright 1942, 1943, Ehrlich & Raven 1969, Lee & Mitchell-Olds 2011, Wang et al. 2013). Next-generation sequencing can detect large numbers of single nucleotide variants (SNVs) genome-wide which can then be inspected to understand evolutionary history between and within species (Gompert et al. 2014). Levels of admixture, shared genomic lineages within an individual, may vary along a geographical gradient (i.e. IBD) or by ecological boundaries (i.e. IBE), which can give insight into the mechanism restricting gene flow.

Here, I test for genomic patterns of IBE among populations of the gall forming wasp, *Belonocnema treatae* (Hymenoptera: Cynipidae), which exhibits regional host plant specialization. Cynipid wasps are among the most specialized herbivorous insects as they only feed and develop on a single or series of closely related host plants (Askew 1984, Quicke 1997). The highest diversity of Cynipids are found on oaks belonging to the genus *Quercus* (Askew 1984, Stone et al. 2002). *B. treatae* induce galls on three different species of live oaks (*Quercus,* subsection *Virentes*) found in the southern USA: *Q virginiana* (*Qv*), *Q. geminata* (*Qg*), and *Q. fusiformis* (*Qf*) (Mayr 1881, Muller 1961). In regions of sympatry of the sister species *Qv* and *Qg* (Nixon 1985, Cavender-Bares &

Pahlich 2009) in the southeastern US, *B. treatae* exhibits partial positive assortative mating, host-associated habitat preference and differences in adult and gall morphology (Egan et al. 2012a, 2012b, 2013). Given this pattern of host associated behavioral and morphological differences in *B. treatae,* I hypothesize that each oak species is acting as a different environment promoting divergent selection within *B. treatae.*

Live oaks belonging to the *Virentes* subset of *Quercus*, have a long complex evolutionary history driven by geological and climatic changes (Nixon 1985, Manos et al. 1999, Cavender-Bares et al. 2004, 2011, 2015, Cavender-Bares & Pahlich 2009, Pearse & Hipp 2009, Gugger & Cavender-Bares 2013). *Virentes* represent oak species that are extraordinarily woody, have high levels of tannins and retain their green canopy throughout winter in the southeastern US and Texas, with a typical leaf lifetime of one year (Nixon 1985, Cavender-Bares & Holbrook 2001). Within *Virentes* there are seven species that vary in range size and contact with other species: *Q. brandegeei* (*Qb*), *Qf*, *Qg*, *Q. minima* (*Qm*), *Q. oleoides* (*Qo*), *Q. sagraeana* (*Qs*) and *Qv*. *Virentes* are estimated to have a crown age of 11 million years (8.4-14.1 MA) based on fossil calibration; consisting of two clades in the US: 1) western clade with *Qf* and 2) eastern clade with *Qv* and *Qg* (Cavender-Bares et al. 2015). Introgression within *Virentes* is not uncommon, especially in areas where *Qf* coexist with other species, such as at the eastern *Qv* range or the northern *Qo* range near Mexico (Muller 1961, Cavender-Bares et al. 2015). All members of *Virentes* are wind pollinated and interfertile (Nixon 1985), but are typically reproductively isolated due to differences in flowering time and niche occupation (Cavender-Bares & Pahlich 2009). *Virentes* are typically found in habitats with well-drained sandy soil or volcanic tuff (Cavender-Bares et al. 2004); each species occupies

their own niche: (1) *Qf* is found at higher elevations and drier soils than *Qv* and (2) *Qv* is found in wetter nutrient rich habitats compared to the drought resistant *Qg* (Cavender-Bares & Pahlick 2009).

Cavender-Bares et al. (2015) investigated the evolutionary relationship among all members of *Virentes* using 3 molecular techniques: phylogenetics using RADseq data, genetic structure analysis using 11 nuclear single sequence repeats (SSRs), and further quantified genomic diversity with chloroplast haplotype networks using sequences within *trn*D-*trn*T and *rpl*32-*trn*L^{UAG} regions. Their findings show the largest division among the seven species occurs at $k = 2$, separating *Qf* and *Qb* from the other five southeastern species, although they did not focus on this division. All three molecular analyses show signs of two distinct genetic lineages with varying degrees of substructure within the lineages. The lack of association between *Qf* and *Qv* in their analysis is likely a reflection of their sampling design, as no samples were included from regions known to have hybrids (such as *Qf* and *Qv* in eastern Texas). This presents a large gap in their sampling design, but allows for easier phylogenetic analysis. Analyses of these putatively "pure" species, show signs of *Qf* and *Qb* diverging from the eastern lineages (Florida's: *Qv, Qg* and *Qm*, Mexico's *Qo* and Cuba's *Qs*) around the same time *Qv* diverged from *Qg*, ~8 Ma (Cavender-Bares et al. 2015). However, this contradicts previous predictions based on phenotypic characteristics that interpreted *Qf* diverging from *Qv* and *Qo* lineages (Nixon 1985). While occupying the same geographical regions, *Qv* and *Qg* are believed to have diverged during periods of island barrier fluctuations around 8 Ma in the Florida peninsula (Cavender-Bares et al. 2015). Unsurprisingly, genetic diversity within species was found to increase as range size increased, with *Qv* and *Qo* having similarly high

levels of diversity followed by *Qf*. *Qf's* range is believed to have been thirty times larger than it is today, but has contracted due to climatic and drying changes; likely contributed to the relatively high genetic diversity seen in SSR and chloroplast estimates of diversity (Cavender-Bares et al. 2015).

Spatial variation is inevitably part of any ecological study, especially at the large spatial scale investigated in this study, as organisms have limited dispersal abilities and may be restricted by geographic barriers. This can propagate spatial autocorrelation, making it difficult to disentangle evolutionary processes from spatial structure and IBD (Legendre & Fortin 1989, Legendre et al. 2015, Radersma et al. 2017). The Mantel test (Mantel, 1967) has commonly been used by ecologists to test for spatial autocorrelation with environmental effects. Unfortunately, in many cases the Mantel test has been misused and is inappropriate when combining spatial data with ecological data (Legendre et al. 2015). Simulations have demonstrated there is autocorrelation involved in creating the dissimilarity matrices, resulting in violations of the assumptions of linearity and homoscedasticity (Legendre & Fortin 2010, Legendre et al. 2015). Legendre et al. (2015) proposed a more appropriate approach to detect signals of spatial structuring using distance based Moran's eigenvector mapping (dbMEM), formally known as principal coordinates of neighbor matrices (PCNM) (Borcard and Legendre 2002, Borcard et al. 2004, Dray et al. 2006). To account for spatial autocorrelation associated with the distribution of sample sites within the sampling design landscape, dbMEM are used to construct regression variables summarizing spatial structure. These explanatory spatial variables can then be used in canonical analysis, such as Redundancy analysis (RDA),

along with other environmental predictors to identify the proportions of variation due to spatial and environmental factors (Borcard et al. 2004).

In this study, I sampled populations of *B. treatae* throughout the geographic range of *Qf, Qv* and *Qg* across the southern US. I had two objectives: (1) assess whether live oak species are acting as a source of divergent natural selection in *B. treatae,* as inferred by patterns of genetic differentiation among host plant affiliated gall former populations and if so, ascertaining whether the patterns of genomic divergence within *B. treatae* parallels the phylogeography of the host plants and (2) partition genomic variation due to spatial autocorrelation to better understand how much genomic variation is attributed to host association versus geography. To investigate the role of host plant affiliation in determining patterns of among population genetic differentiation of a regional host plant specialist, I implemented an RDA analysis using dbMEM spatial variables to partition genomic variation among insect populations into host plant affiliation and spatial landscape components. This is important because even if patterns of *B. treata*e's genomic variation parallels host plant phylogeography, the observed patterns may be due to the spatial distribution of live oak species rather than divergent natural selection. Thus, the dbMEM approach provided a reliable way to test the significance of host plant in driving patterns of genomic variation within *B. treatae,* while controlling for spatial effects.

II. MATERIALS AND METHODS

Study System and Host Plants of Belonocnema treatae

Belonocnema treatae (Hymenoptera: Cynipidae) (Mayr 1881) is widely distributed across the geographic range of American live oaks subseries *Virentes* (Muller 1961, Cavender-Bares et al. 2004, 2011, 2015, Cavender-Bares & Pahlich 2009, Pearse & Hipp 2009, Gugger & Cavender-Bares 2013) across the southern and southeastern US, (Lund et al. 1998, Egan et al. 2012a, 2012b, 2013). Similar to many cynipids (Stone et al. 2002), *B. treatae* undergo an alternating life cycle with temporally distinct sexual and asexual generations, in which adults are restricted to inducing galls on newly developing tissue of the leaves and roots, respectively (Lund et al. 1998). Galls are 3-dimensional structures found on host plants that are induced following oviposition onto plant tissue by gall-forming insects (Malyshev 1968, Askew 1984, Tooker and De Moraes 2008). Upon successful evasion of the host immune system, the growth, development and nutritive value of the gall is controlled by the insect (Malyshev 1968, Askew 1984, Price et al. 1987, Tooker and De Moraes 2008). Emergence of the sexual generation coincides with the leaf flush of the host during spring (Hood and Ott 2010), as this is when the leaves are most susceptible to gall induction because the leaf cuticle has not fully developed. The sexual generation mates and oviposit unto the lateral veins on the underside of leaves throughout the tree crown, giving rise to single chambered galls within which a single asexual female develops. The asexual generation emerge from October – December and oviposit on newly developing root tissue, producing multi-chambered galls that give rise to all males or all female broods. Gall traits are an important component of fitness, as galls act as microclimates (Price et al. 1987, Miller et al. 2009) that provide the insect

their sole source of nutrients, as well as some protection from environmental conditions and natural enemies (Stone et al. 2002). Leaf galls induced by *B. treatae* can vary from 0.5 -9.0 mm in diameter and gall size is positively correlated with adult body size (Lund et al. 1998, Egan et al. 2013), this in turn translates into increased fecundity (Hood & Ott 2017). Gall phenotypes are a product of the interaction between plant and insect genotypes (Stone et al. 2002), thus neighboring live oak trees may differ greatly in their quality for gall induction and gall size (Egan & Ott 2007) as well as *B. treatae* fitness (Egan et al. 2011).

Historically, cynipids are thought to have radiated from Central America along with the geographical range shifts of their oak hosts (Manos et al. 1999, Stone et al. 2002). Members of *Virentes* are widely distributed across southeastern US, with varying degrees of species overlap. *Qf* spans from central Texas southward into Mexico with isolated populations in both southwestern Oklahoma and along the western edge of the Edwards Plateau in Texas. *Qv* is distributed from eastern Texas across the Gulf Coast into Florida and along Atlantic coast to extreme coastal southern Virginia. *Qg* has the smallest range being mainly found in Florida, with isolated populations spanning to Alabama (Nixon 1985, Manos et al. 1999, Cavender-Bares et al. 2015). While *Qv* and *Qg* typically occupy different niches within the same geographical range, they can co-occur within meters of each other (Cavender-Bares & Pahlich 2009). *Virentes* are distinguished by phenotypic variation - leaf morphology and flowering time (Cavender-Bares et al. 2004, 2011, 2015, Cavender-Bares 2007, Cavender-Bares & Pahlich 2009) that may be able to promote plant-driven diversification of herbivorous insects (Funk et al. 2002) Ecologically important traits (gall size and body size) vary among host-associated

populations of *B. treatae* inhabiting *Qg* and *Qv* in Florida, indicating that divergent natural selection is likely occurring (Egan et al. 2012a, 2013). As well, host associated populations of *B. treatae* on *Qv* and *Qg* in Florida display strong mating and oviposition preferences for native host species (Egan et al. 2012a, 2012b) which act as reproductive barriers by reducing mating encounters among individuals associated with *Qv* and *Qg* hosts. Field and greenhouse experiments with native *Qf* associated *B. treatae* demonstrated immigrant inviability, as *B. treatae* were less successful at inducing galls on novel *Qg* hosts (Zhang et al. 2017). Such patterns may be driven by the leaf structure of the plants: *Qg* have significantly more trichomes and thicker leaves compared to *Qv*. This difference in leaf structure may require *B. treatae* to adapt differently to deal with the tough oviposition and gall development conditions *Qg* pose. Another possible adaptive difference is *B. treatae* associated with *Qg* consistently have shorter wings than *Qf* and *Qv* insects and do not appear to fly (personal observation). Moreover, genome size analysis of *B. treatae* has revealed significant differences in genome size among Texas (*Qf*) and Florida (*Qv* and *Qg*) populations (Hjelmen et al. 2013), providing preliminary support of host-associated genetic divergence.

Inspection of neutral mitochondrial markers (cytochrome *b* and cytochrome oxidase I (COI)) of *B. treatae* collected from two sympatric hosts populations in Florida (*Qv* and *Qg*) showed no sign of variation due to host affiliation (Egan et al. 2012a). A nuclear marker was also tested, the 257bp ITS2 region, but no variation was observed consistent with recently diverging lineages (Egan et al 2012a). However, two major geographically distinct mitochondrial haplotype clades were identified by genotyping COI on for *B. treatae* individuals sampled from 23 populations distributed across the

geographic range of the three host plants (Schuler et al. 2018). The two clades identified followed a geographical pattern where a western clade spanned Oklahoma/Texas to Mississippi (*Qf* and *Qv* sites) and an eastern clade spanned from Mississippi to Florida (*Qv* and *Qg* sites); with individuals from Gautier, MS belonging to both clades (Schuler et al. 2018). Each of these clades were also associated with different strains of the endosymbiont *Wolbachia*, with Gautier having individuals with *Wolbachia* strains associated with the different haplotype clades (Schuler et al. 2018). This finding is salient with respect to inspection of genetic divergence among host plant affiliated *B. treatae* populations as the possibility of reproductive isolation in the form of cytoplasmic incompatibility (CI) is among the numerous phenotypic effects that *Wolbachia* can have on their hosts (Werren 1997). CI occurs when individuals that are infected with different stains mate, but are unable to produce offspring due to incompatibles of the sperm and egg during fertilization caused by *Wolbachia* (Werren 1997). *Wolbachia* induced selective sweeps can occur when a *Wolbachia* infection quickly spreads through a population and carries along the mitochondrial haplotypes of the few infected founders, consequently reducing mitochondrial diversity (Turelli & Hoffmann 1991, Turelli et al. 1992). Thus, it is possible that this maternally inherited bacterium may be responsible for driving or maintaining the haplotype pattern observed. This necessitates the need for a genomic analysis, which is robust to patterns of divergence and genetic diversity caused by *Wolbachia* induced selective sweeps at the mitochondrial level.

Sampling and Data Collection

Samples were collected just prior to the onset of the maturation and emergence of the asexual generation *B. treatae,* beginning in mid-October thru late November of 2015 and 2016. In 2015 I collected from 50 sites and 68 sites in 2016, with some sites were visited in both years to augment sample sizes. Individuals from the same site across years were pooled together as there is no reason to believe it would affect population genomic structure. Collection localities span the geographic ranges of the three live oak species across the southern US, including an isolated population in Oklahoma and western edge of the Edwards Plateau in central Texas (Figure 1; Table 1). On the Atlantic coast, I surveyed the northern range limits of *Qv,* finding the northern range limit of *B. treatae* to be at Beaufort NC, despite an exhaustive search of *Qv* populations extending hundreds of kilometers northward throughout coastal North Carolina and southern Virginia. I did not sample *B. treatae* from any oaks classified as *Qm*, because they are not common and are likely phenotypically plastic form of *Qg* as there is little molecular support to classify *Qm* as a separate species (Cavender-Bares et al. 2015). Leaf gall abundance varies dramatically among trees within populations of live oaks (Egan et al. 2007) and occurrence of *B. treatae* can be patchily distributed at the landscape level. However, oak trees populated by moderate to high densities of *B. treatae,* are however readily detected, thus to locate sample sites I drove public roadways stopping at intervals to inspect individual, clumps, or scores of live oak trees. In total, I collected 126,812 leaf galls (mean = 1093 ± 103.32; median = 741.5 per site) that subsequently produced 9,000+ asexual *B. treatae.*

Live oak species at each collection locality were identified based on morphology and range distribution as described by Cavender-Bares et al. (2015). To reduce the chance of including siblings in genomic analyses, leaf galls were collected by widely searching each tree to reduce resampling of high density clusters of galls within trees that could represent a single female's oviposition efforts and by distributing sampling effort across multiple trees/site where possible. Collected leaf galls were stripped from the leaves, pooled by site and housed in collection traps maintained outdoors at the Texas State University Research Greenhouse in a shaded alcove under natural weather conditions. Collection traps were monitored daily for the emergence of adult *B. treatae*. Upon emergence, *B. treatae* were collected alive and stored in 95% ethanol at 4°C until DNA extraction.

In total, 58 of the 94 collection localities were used in this analysis (~30 individuals per site) and represented a broad sampling across *B. treatae's* geographical range. In total the analysis included (N = 14 *Qf*, N = 36 *Qv* and N = 8 *Qg* sample locations). At one site (Amelia Island, FL) the *Qv* and *Qg* host plant species were sympatric, with *Qv* and *Qg* trees occurring within meters of each other. The *B. treatae* collected from each tree species at this site are treated as two *B. treatae* samples (each with N = 30 *B. treatae*) based on host affiliation, but with the same GPS coordinates (sites 47 and 48). One site (Gulf Shores, AL), had sample sizes supplemented with female *B. treatae* from the sexual generation root galls (N = 9) collected spring 2016. Each root gall was isolated and reared in standard fruit fly vials. To prevent sampling of siblings, only one female per root gall was used. I sequenced both generations at Gulf Shores because females from both generations are diploid and are not believed to vary in

genome size (Gokhman et al. 2015). I found no significant difference in individual sequence coverage between the two generations, so the sexual females were included in our analyses.

Genomic Library Preparation:

DNA was isolated from a subset of *B. treatae* reared from leaf galls: 1,536 individuals (average 26.5 ± 6.1 , median = 22 individuals per locality; minimum 7 and maximum 30) by grinding up the entire adult *B. treatae* and following the DNeasy Blood and Tissue Kit and protocol (Qiagen Inc.). I created a reduced representation genomic library for each individual using a highly multiplexed genotyping-by-sequencing (GBS) approach, following the protocols of Parchman et al. (2012), Gompert et al. (2014) and Mandeville et al. (2015). Briefly, genomic DNA was digested using two restriction enzymes, *Eco*R1 and *Mse*1, at non-targeted sites throughout the genome. Customized Illumina adaptor sequences containing the primer sequences and unique 8-10 base pair barcode (individual identifiers) were ligated to DNA fragments. Two separate PCR's were performed on each sample and pooled PCR products were size selected for fragments 250-350 base pairs in length using BluePippin quantitative electrophoresis at University of Texas Genomic Sequencing and Analysis Facility (Austin, TX, US). Before and after size selection, DNA concentration and quality were verified using a BioAnalyzer prior to Illumina sequencing (University of Texas, Austin). Each pooled genomic library was sequenced on two lanes of Illumina HiSeq 4000 platform, with single-end reads of 150bp, for a total of 4 lanes.

Assembly and Variant Calling

For each sequence read, I used a custom perl script to parse restriction sites, correct barcode IDs that were off by one base due to sequencing errors and replace sequence barcode IDs with individual IDs for indexing. I removed sequences that had short reads or that contained *Mse*1 adaptor sequences. An artificial reference genome was then created using the dDocent *de novo* assembly based on the contigs generated after parsing the sequence data of all individuals. I specified that candidate sequences for inclusion into the reference genome must have 4 or more reads to be selected and be represented by ≥ 4 individuals. Further, CD-HIT (Cluster Database at High identity with Tolerance) was then performed with a minimum of 80% similarity. Consensus sequences from this *de novo* assembly were used as an artificial reference. All sequences were then assembled to the reference using BWA ver. 0.7.13 (Burrows-Wheeler Aligner, Li & Durbin 2009).

To identify bi-allelic single nucleotide variants (SNVs), I used *samtools* ver. 0.1.19 and *bcftools* ver. 0.1.19 (Li et al. 2009). Variants were called when 90% of all individuals had at least one read at a given SNV. I used a full prior for variant calling and set the threshold probability for identifying a variant site at $P = 0.05$. I incorporated genotype uncertainty due to sequencing and alignment errors in downstream analysis by retaining genotype likelihoods (Li 2011, Skotte et al. 2013). After I identified SNVs, the SNVs with more than one alternative allele were removed (to eliminate potential paralogs). From each contig I selected a single SNV reducing potential linkage disequilibrium between SNVs. I then sorted variants by minor allele frequencies (MAF),

retaining SNVs with a MAF >5%. I removed low coverage individuals (<1x median coverage), then repeated subsequent filtering starting at variant calling.

Population Genetic Structure

Population genomic structure was estimated using the program *entropy* (Gompert et al. 2014). This hierarchical clustering Bayesian model incorporates uncertainty in sequencing coverage and error within loci and estimates of allele frequencies, and calculates genotype probabilities based on estimated genotype likelihoods (Gompert & Buerkle 2010, Gompert et al. 2010). *Entropy* only requires a specification of the number of ancestral clusters (*k*) without the need for a priori assumptions about an individual's assignment probability. *Entropy* produces estimates of genotype probabilities, genomic clusters and admixture proportions (*q*) of *B. treatae* individuals. I ran *entropy* models $k = 2$ to $k = 10$ to capture population genomic structure. Posterior estimates of genotype probabilities were obtained by running 75,000 Markov Chain Monte Carlo (MCMC) steps with a 5,000 step burn in and thinning by retaining every 10^{th} value. MCMC mixing and convergence were checked by estimating effective sample size (ESS) using *coda* in R (Plummer et al. 2006, R Core Team 2015) and examining Gelman-Rubin diagnostics, respectively. ESS is a measure of model mixing, each ESS value represents the number of independent MCMC steps, thus higher ESS scores indicate less autocorrelation between steps. Gelman-Rubin diagnostics evaluates chain convergence, values between 1.0 and 1.1 are optimal as they indicate the chains for each *k* are arriving at the same conclusion (Gelman & Rubin 1992). Mean assignment probabilities (*q*) were averaged between the two chains run for each *k* model. Mean genotype posterior probabilities were

checked for correlations, then averaged between chains and across *k*'s. PCA was performed on the genotype posterior probabilities to visualize the patterns of genomic differentiation among populations.

Dividing Data by Eastern and Western localities

To visualize patterns of genomic variation at finer scales, following the above analyses, I subset the data by classifying the individual *B. treatae* according by the pattern of genomic division seen at $k = 2$ (corresponding to east and west localities, Figure 2). This division also corresponds with the haplotype clades seen using COI marker (Schuler et al. 2018). The West subset is comprised of 593 individuals (sites 1 – 30) representing samples collected from host plants identified as either *Qf* or *Qv*. The East subset is comprised of 651 individuals (sites 30 – 58) representing samples collected from *Qv* and *Qg* host plants. Individuals from site 30 (Gautier, MS) were included in both subsets because this location is known to have individuals assigning to two different lineages. SNVs of each subset were processed the same as the larger datasets, but variant calling was restricted at 50%. Proportions of admixture were obtained using *entropy*, as described above.

Genetic Distance and Diversity

To compare the levels of genomic variation within and among collection localities and with respect to host plant affiliation, mean genotype probabilities were used to calculate allele frequencies. G_{ST} values (Nei 1973) were computed to estimate genetic distance within a subpopulation in relation to the total genetic variance based on allele

frequencies. To visualize among-population level genetic variation, Nei's D_A was computed and used to construct a dendrogram using *ape* ver. 5.0 in R (Nei et al. 1983, Takezaki & Nei 1996). Population level genetic diversity was calculated using the expectation-maximization (EM) algorithm with 20 iterations for the Watterson's θ (number of segregating sites-SNVs) and Tajima's π (nucleotide diversity) with *samtools* and *bcftools* (Li et al. 2009, Li 2001).

Partitioning the roles host and spatial variance

To partition genomic variation due to the spatial autocorrelation and ecological processes (host plant), I incorporated distance-based Moran's eigenvector mapping (dbMEMs, formally called PCNM) into a Redundancy analysis (RDA) (Borcard and Legendre 2002, Borcard et al. 2004, Dray et al. 2006). When comparing matrices of spatial data and genetic distance data, dbMEM has higher power and is more appropriate than the Mantel test (Legendre & Fortin 2010, Legendre et al. 2015). This stems from the autocorrelation involved in creating the dissimilarity matrices used in the Mantel test, resulting in violations of the assumptions of linearity and homoscedasticity (Legendre & Fortin 2010, Legendre et al. 2015).

Spatial predictor variables were extracted as dbMEM variables following Borcard & Legendre (2002). To construct dbMEM variables, I first, constructed a pairwise geographic distance matrix using haversine (great-circle) distances between all pairs of locations from which all of the 1,219 individuals retained in the analysis were sampled, using the package *geosphere* ver. 1.5-5 in R (Nychka et al. 2017). Because the RDA involves the genomic data of individuals, all latitude/longitude coordinates were

considered on an individual level. Thus the pairwise distance matrix contained 2^{58} pairs of distances. Next, a minimum distance spanning tree (Borcard et al. 2002) was created for the 58 sites to identify the maximum nearest-neighbor distance (minimum spanning distance) across all pairs of sites. This value (433km), multiplied by four was then used to replace all pairwise Euclidean distances that exceeded this threshold near neighbor distance of 1,732 km (Borcard et al. 2004). This truncated distance matrix allows all individuals in comparison to all other individuals to be binned into categories of "neighbors" (< threshold) or "not neighbors" (> threshold). Due to the complexity of the data and to reduce computational time, I collapsed the genotype probabilities matrix using the *vegdist* function in R package *vegan* ver. 2.4-4. This approach collapses the columns (loci) of the genotype probability matrix by Euclidean space to give a composite genotype probability for each individual. Second, dbMEM was calculated in *adespatial* ver. 0.1-0 in R on the collapsed genotype probability matrix using coordinates as the predictors. DbMEM implements a principal coordinate analysis (PCoA) on the truncated Euclidean distance matrix with the collapsed genotype probability matrix. The positive eigenvectors were then selected as spatial variables (dbMEMs) as they are positively correlated with distance. Third, a distance based RDA (dbRDA) was performed using the positive axes as spatial predictors and host affiliation as an environmental predictor. The resulting residuals then underwent a permutation ANOVA to identify the significant axes. Finally, dbRDA on the collapsed genotype probability matrix was used with the significant positive dbMEMs as spatial predictors and host plant affiliation of each *B. treatae* sampled as the environmental predictor. Significance of the dbRDA predictors

were assessed again using a permutation ANOVA to determine the proportion of genomic variation explained by host plant.

III. RESULTS

Assembly and Variant Calling

After parsing the sequence reads from the 1,536 individual *B. treatae*, a total of 964,115,687 parsed reads were retained, of which 158,978 contigs met the criteria to be used in the construction of the reference genome. I next removed 317 individuals with a median coverage of ≤1x, this winnowed data set of 1,219 individuals was used henceforth. Variant calling identified 6,986 loci, with an average median sequence coverage of 5.54 and an average mean of 11.47 reads per individual.

Population Genetic Structure

Entropy identified a strong east/west geographic division at Gautier, MS which is maintained through *k*2 – *k*10 (Figure 2). At *k* = 3 a third cluster is resolved within the eastern cluster corresponding to *B. treatae* collected from *Qg* host plants as opposed to *Qv* hosts. At *k* = 4 evidence of mixed ancestry in the western division appears, especially within *Qf* associated *B. treatae*. At *k* = 5 sites 30 – 35 corresponding to the eastern Gulf Coast form a cluster showing signs of IBD in eastern *Qv* host sites. At *k* = 6 the admixture seen at *k* = 4 retreats to just western *Qv* sites and is replaced in *Qf* sites with a new genetic cluster (light blue). Beyond *k* = 6, *entropy* models start breaking down as evidenced by high Gelman-Rubin scores and small EES value. Models beyond *k* = 6 were thus deemed not reliable (Table 2).

The PCA on mean genotype probabilities formed three clusters, corresponding with *k* = 3 structure (Figure 3). PC1 accounted for 68.6% of variation, strongly separating the eastern and western localities. PC2 (23.5%) separated B. *treatae* associated with *Q.*

geminata while PC3 (1.5%) weakly separates the western localities, predominantly by host plant (*Qf* and *Qv*).

Eastern and Western Subset Structure

East and West subset data was used to simplify the *entropy* models to improve model performance by removing individuals that are very diverged (i.e. east/west division). The East subset (sites 1 to 30) data performed well at *k*2 – *k6*. Overall the Eastern loci showed the same pattern as the study wide data, even thru $k = 10$ which were initially deemed not reliable, but with higher resolution and better model performance (Figure 4). At $k = 2$, eastern *Qv* affiliated *B. treatae* populations are separated from those collected at *Qg* sites, with this division maintained throughout all *k*s. At $k = 3$ the same eastern Gulf Coast clustering from the larger dataset at $k = 5$ is evident at sites 30-35. At $k = 4$, sites 51 and 55 start clustering together. At $k = 5$, *Qg* populations are separated by their geographic ranges, with the northwestern Florida panhandle sites (36, 37, 38, 40) clustering independent from the eastern range (48, 53,54,56). At $k = 6$ the model starts breaking down, as seen by the 50/50 assignment probabilities within *Qv* sites 41-58.

The West subset data performed well at *k*2 – *k6*. At $k = 2$, the same clinal admixture seen in *Qf* sites beginning at $k = 4$ of the study wide dataset is present, with the majority of *Qv* affiliated individuals showing no admixture (Figure 5). At $k = 3$ however, *Qf* individuals show evidence of two admixed lineages; *Qv* individuals found at their western most range limits share admixture with *Qf* individuals. At $k = 4$ the individuals from Gautier, MS (site 30) that assign to the eastern divide cluster. At $k = 5$ the geographically isolated Oklahoma site (1) shows complete assignment to its own cluster,

with evidence of their linage in other *Qf* associated *B. treatae*. At $k = 6$ the model becomes less informative as the new cluster does not have 100% assignment to any individuals, but this pattern is also observed in the study wide data set for $k8 - k10$.

Estimates of Genetic Distance and Diversity

Calculations of site-level genomic variation using G_{ST} and genetic distance using Nei's D_A, form three genetic groups that corresponding to the genomic structure seen at $k = 3$ (Figure 6). G_{ST} estimates of genetic distance indicate large differences between each pairwise host plant comparison. *B. treatae* from the geographically isolated *Qf* and *Qg* sites had the largest genetic distance (0.766 $\pm$0.007 CI); followed by the co-occurring *Qv* and *Qg* sites (0.450 $\pm$0.008 CI), then the parapatric *Qf* and *Qv* sites (0.289 $\pm$0.004 CI).

Genome-wide measures of Watterson's θ showed that genetic diversity was highest in *Qg* sites, followed by the eastern *Qv*, then by the western division sharing similar levels of SNVs (Figure 7). Nucleotide diversity was highly variable within host associated *Qv* and *Qg* sites and inconsistent with estimates of Watterson's θ, suggesting a lack of genomic neutrality. However, *Qf* sites demonstrate consistently higher π than the other host associations and had similar scores of genetic diversity as Watterson's θ, consistent with genomic neutrality.

Partitioning of Spatial Variance – dbMEM

To determine the role of spatial autocorrelation and host association on the genomic variance within *B. treatae*, dbMEM analysis was first conducted on spatial coordinates and identified 13 positive axes. Preliminary dbRDA on the collapsed

genotype probability matrix using only latitude and longitude as predictors explained 24.0% of the genomic variation. Permutation ANOVA on spatial dbRDA identified 11 of 13 MEM axes as significant contributing to genomic variation. DbMEM analysis on collapsed genotype probabilities using host affiliation and 11 significant spatial MEM axes as predictors explained 23.3% of the genomic variation. Permutation ANOVA identified host and all 11 spatial MEM variables as significant. Within *B. treatae*, host plant affiliation explained 7.1% of the constrained variation and 1.66% of the total genomic variance. Spatial components explained 92.88% of the constrained variation and 21.64% of the total genomic variation. RDA axis 1 explained 93.3% of the variation and was loaded the heaviest on the spatial variable MEM1 (Figure 8; Table 3). RDA axis 2 explained 2.9% of the variation and was loaded the heaviest on host association.

IV. DISCUSSION

In this study I examined patterns of genomic variation of a widely distributed but regionally host specific gall-former in the southern US and asked whether spatial patterns of genomic variation among sample sites were associated with host plant affiliation and spatial autocorrelation within the sample design. I found both geography and host plant association were significant in structuring patterns of genomic differentiation in *B. treatae*. As predicted by mitochondrial haplotype clades (Schuler et al. 2018), geographic isolation separated collection sites east and west of coastal Mississippi, independent of host association. Within the eastern sites, there is no evidence of gene flow between sympatric *Qv* and *Qg* host associated *B. treatae*, suggesting these hosts harbor different distinct lineages of *B. treatae*. Conversely, western sites illustrate varying degrees of gene flow between parapatric *Qf* and *Qv* host associated *B. treatae* as expected with host associated lineages coupled with IBD.

Consistent with previous studies that suggested oak species are acting as a source of divergent natural selection (i.e. partial positive reproductive isolation and differences in fitness traits), I found patterns of genomic variation in *B. treatae* that paralleled host plant association, particularly in eastern sites. Host associated patterns of genomic structure in *B. treatae* suggest that the sympatric *Qv* and *Qg* hosts are acting as different environments promoting divergent natural selection. The lack of gene flow between these host associated *B. treatae* is particularly convincing given that sympatric Amelia Island, FL (site: 47 and 48), where *B. treatae* were collected from both *Qv* and *Qg*, show no evidence of shared lineages among any of the individuals examined in the *entropy* models (other than at $k = 2$) despite the host associated populations co-occurring within

the dispersal distance of *B. treatae*. While these findings suggest *Qv* and *Qg* host associated *B. treatae* are reproductively isolated, previous studies have shown *B. treatae* from different hosts are capable of producing viable offspring, which begs the question: What maintains the *Qv* and *Qg* host associated reproductive isolation in *B. treatae?* Future studies should explore this isolation more thoroughly to understand if it is driven by intrinsic factors (i.e. reduced hybrid fitness) or extrinsic factors (i.e. host plant).

Western sites reflect a more complex evolutionary history with multiple lineages undergoing admixture coupled with IBD between parapatric *Qf* and *Qv* host associated *B. treatae.* Broadly, western *Qv* host associated *B. treatae* assign to their own lineage with evidence of admixture between neighboring *Qf* sites. Luling, TX represent an exception to this general pattern with all individuals consistently assigning to *Qf* associated lineages. The *Qv* oaks at Luling represent a population of planted trees in a region that is predominantly *Qf* oaks, which is likely the reason why all the *B. treatae* assigned to *Qf* lineages. Within *Qf* host associated *B. treatae,* there appear to be two distinct lineages with varying degrees of admixture. One *Qf* host associated lineage primarily occurs in central Texas, while the other is prominent along coastal Texas. These two *Qf* host associated lineages in *B. treatae* geographically co-occur where there is evidence of multiple lineages in the *Qf* oaks (Cavender-Bares et al. 2015). Historically, *Qf* oaks had a range thirty times larger than it is today, neighboring and exchanging genes with *Qb, Qo* and *Qv*. Given these patterns of multiple lineages in *Qf* oaks and *Qf* associated *B. treatae*, and the strong east/west division that separates *Qv* associated *B. treatae,* it may be more appropriate to view this system as having 4 or 5 oak hosts rather than 3 (*Qf, Qv* & *Qg*). There is longstanding evidence that specialized herbivorous insects can be used to infer

the evolutionary history of their host (Ehrlich & Raven 1969, Hafner & Nadler 1988, Mitter et al. 1991). Because herbivorous insects are typically restricted to feeding upon a single or closely related host species, they share a close evolutionary history with their host (Ehrlich & Raven 1969, Funk et al. 2002). Depending on the history of the extent to which the insect lineage(s) have been associated with the host lineage(s), there may be patterns of coevolution (Mitter et al. 1991). Therefore, given the patterns seen here, it is possible I am underestimating of the effect of host plant association as a predictor of genomic variation in *B. treatae* when categorizing hosts into only three species.

This study provides support for host driven patterns of divergence, as expected with IBE. Given the strength of the genomic distance and variation of *Qg* affiliated sites compared to *Qv* affiliated sites suggest *Qg B. treatae* have a long history of reduced gene flow with the sympatric *Qv* affiliated *B. treatae*. This reproductive isolation in *B. treatae* may stem from *Qv* and *Qg* oaks having temporally isolating flowering time, differences in leaf structure, coupled with *B. treatae*'s partial positive assortative mating preference. While understanding what is maintaining the lack of gene flow between sympatric host associated *B. treatae* is important, from an experimental evolution perspective, testing for ongoing ecological divergence in the face of gene flow may be too late for eastern populations of *B. treatae* affiliated with *Qv* and *Qg* host plants. However, the patterns of admixture between *B. treatae* populations affiliated with *Qf* and western *Qv* host plant among sites suggest there may be ongoing divergence or ongoing homogenization between the lineages. Herein, I demonstrated geography and host plant association are important factors in structuring patterns of genomic variation in *B. treatae,* with varying degrees of influence across the geographic range.

TABLES

Table 1: Sampling localities of *Belonocnema treatae* in southeastern US. Locality information and sample sizes for pre (1,536 individuals) and post (1,219 individuals) filtering of low coverage individuals.
* Gulf Shores, AL used female samples from both generations (asexual N=5, sexual N=9)

Site	Site Number Abbrv.	Host Plant	Latitude (°N)	Longitude (°W)	Initial No. of Inds N = 1536	No. of Inds. Analyzed N = 1219
Quartz Mts, OK	1	*Qf*	34.89	-99.30	30	23
Irion County, TX	2	*Qf*	31.21	-100.84	30	16
Mason County, TX	3	*Qf*	30.82	-99.37	30	30
Schleiker, TX	4	*Qf*	30.90	-100.58	18	18
Liberty Hill, TX	5	*Qf*	30.67	-97.93	30	20
Rocksprings, TX	6	*Qf*	29.88	-100.11	30	30
San Marcos, TX	7	*Qf*	29.94	-98.01	30	23
Luling, TX	8	*Qv*	29.67	-97.63	30	27
Altair, TX	9	*Qf*	29.56	-96.50	30	25
Pleasanton, TX	10	*Qf*	28.95	-98.45	30	25
Rice, TX	11	*Qv*	29.72	-95.40	30	30
Inez, TX	12	*Qf*	28.89	-96.82	30	29
Vidor, TX	13	*Qv*	30.10	-93.93	12	8
High Island, TX	14	*Qv*	29.56	-94.39	30	19
Sulpher, LA	15	*Qv*	30.23	-93.36	16	16
Port O'Connor, TX	16	*Qf*	28.45	-96.42	30	26
Live Oak Park, TX	17	*Qf*	27.85	-97.21	30	18
Egan, LA	18	*Qv*	30.24	-92.53	18	10
Oak Grove, LA	19	*Qv*	29.77	-92.98	30	25
Delcambre, LA	20	*Qv*	29.95	-91.96	30	26
Encino, TX	21	*Qf*	26.89	-98.14	30	18
Baton Rouge, LA	22	*Qv*	30.41	-91.18	15	12
Baldwin, LA	23	*Qv*	29.83	-91.57	22	21
Morgan City, LA	24	*Qv*	29.69	-91.31	30	26
McAllen, TX	25	*Qf*	26.22	-98.24	22	14
Picayune, MS	26	*Qv*	30.53	-89.68	30	25
Golden Meadow, LA	27	*Qv*	29.39	-90.27	30	20
Bay Saint Louis, MS	28	*Qv*	30.32	-89.32	30	26
G. Okeefe, MS	29	*Qv*	30.39	-88.87	30	20
Gautier, MS	30	*Qv*	30.38	-88.61	30	25
Grand Bay, AL	31	*Qv*	30.49	-88.34	30	27
Dauphin Island, AL	32	*Qv*	30.25	-88.13	30	25
Gulf Shores, AL*	33	*Qv*	30.26	-87.72	14	9
Pensecola, FL	34	*Qv*	30.50	-87.23	21	21
N. Santa Rosa Isl., FL	35	*Qv*	30.41	-86.74	30	20
Inlet Beach, FL	36	*Qg*	30.27	-86.00	30	22
Parker, FL	37	*Qg*	30.11	-85.60	30	23
Oceanside Village, FL	38	*Qg*	29.95	-85.43	30	24
N. Highland View, FL	39	*Qv*	29.84	-85.32	7	4
Ochlocknee, FL	40	*Qg*	29.96	-84.39	30	23
Perry, FL	41	*Qv*	30.12	-83.59	30	20
Lanark Village, FL	42	*Qv*	29.89	-83.04	13	12
High Springs, FL	43	*Qv*	29.84	-82.63	24	22

Table 1: Sampling localities of *Belonocnema treatae* in southeastern US.
Continued

Progress Park, FL	44	*Qv*	29.78	-82.47	30	27
Jekyll Island, GA	45	*Qv*	31.02	-81.43	30	21
Sapelo Island, GA	46	*Qv*	31.40	-81.28	30	23
Amelia Island, FL	47	*Qv*	30.52	-81.44	30	26
Amelia Island, FL	48	*Qg*	30.52	-81.44	29	20
Charleston, SC	49	*Qv*	32.77	-79.97	30	25
Palm Coast, FL	50	*Qv*	29.60	-81.20	16	8
Oak Hill, FL	51	*Qv*	28.90	-80.85	30	30
Pawley's Island, SC	52	*Qv*	33.51	-79.06	12	11
Lake Lizzie, FL	53	*Qg*	28.23	-81.18	30	25
Archbold, FL	54	*Qg*	27.18	-81.35	17	17
Kissimmee River, FL	55	*Qv*	27.38	-81.10	30	18
Dickinson State Park, FL	56	*Qg*	27.03	-80.11	30	20
Fort Macon, NC	57	*Qv*	34.70	-76.69	30	20
Beaufort, NC	58	*Qv*	34.71	-76.63	30	25

Table 2: Estimates of *entropy* model performance: ESS and Gelman-Rubin values. Model performance was assessed for the study wide *entropy* runs and the east/west subsets. Large values of ESS indicate the MCMC chain (c) is mixing, while small Gelman-Rubin values indicate chain convergence for a given model (*k*).

	Study wide: 1,219 Inds.		***East Subset: 593 Inds.***		***West Subset: 651 Inds.***	
	ESS	**Gelman - Rubin**	**ESS**	**Gelman - Rubin**	**ESS**	**Gelman - Rubin**
k2c0	3794.55	1.02	2973.45	1.03	2048.44	1.01
k2c1	3818.18	-	2944.64	-	2050.26	-
k3c0	3521.04	1.03	1250.87	1.06	1150.11	1.02
k3c1	3490.72	-	1205.87	-	1147.45	-
k4c0	2120.11	1.04	3327.46	1.01	1166.43	1.02
k4c1	2095.62	-	3329.60	-	1175.09	-
k5c0	2228.59	1.04	1249.44	1.04	1095.90	1.03
k5c1	2239.49	-	1271.69	-	1105.27	-
k6c0	2769.75	1.05	2770.71	1.03	745.97	1.02
k6c1	2796.15	-	2826.81	-	712.55	-
k7c0	2828.73	1.06	-	-	-	-
k7c1	2858.33	-	-	-	-	-
k8c0	2331.72	1.05	-	-	-	-
k8c1	2271.93	-	-	-	-	-
k9c0	2169.58	1.62	-	-	-	-
k9c1	2183.00	-	-	-	-	-
k10c0	2188.93	1.07	-	-	-	-
k10c1	2202.01	-	-	-	-	-

Table 3: dbRDA loading scores of spatial autocorrelation and host plant predictors.

	dbRDA1	dbRDA2	dbRDA3	dbRDA4	dbRDA5	dbRDA6	dbRDA7	dbRDA8	dbRDA9	dbRDA10	dbRDA11
host	0.472	0.499	-0.680	-0.054	0.011	0.032	0.179	-0.157	0.000	-0.019	0.064
MEM1	0.869	-0.078	-0.408	-0.018	0.087	0.007	-0.220	-0.051	0.111	0.040	-0.002
MEM2	-0.329	0.032	-0.650	0.613	0.226	-0.039	-0.012	-0.033	0.055	0.122	0.143
MEM3	0.213	-0.187	0.614	0.611	0.187	-0.145	-0.053	-0.135	0.234	0.070	0.184
MEM4	-0.059	-0.404	-0.152	-0.263	0.122	-0.567	0.395	-0.334	0.243	-0.251	0.123
MEM6	0.116	-0.032	-0.036	0.147	-0.336	-0.404	0.210	0.795	0.032	-0.007	0.096
MEM7	0.061	-0.376	-0.054	0.014	-0.450	0.352	0.359	-0.172	-0.127	0.454	0.381
MEM9	0.089	0.048	0.070	-0.195	0.739	0.167	0.404	0.303	-0.162	0.223	0.204
MEM10	0.067	-0.056	0.034	0.087	0.032	-0.468	-0.041	-0.172	-0.717	0.358	-0.301
MEM11	0.185	0.094	-0.011	0.275	-0.066	0.189	0.330	-0.094	-0.476	-0.695	0.116
MEM12	-0.060	0.164	0.021	-0.195	0.001	-0.222	-0.450	-0.028	-0.224	-0.056	0.791

FIGURES

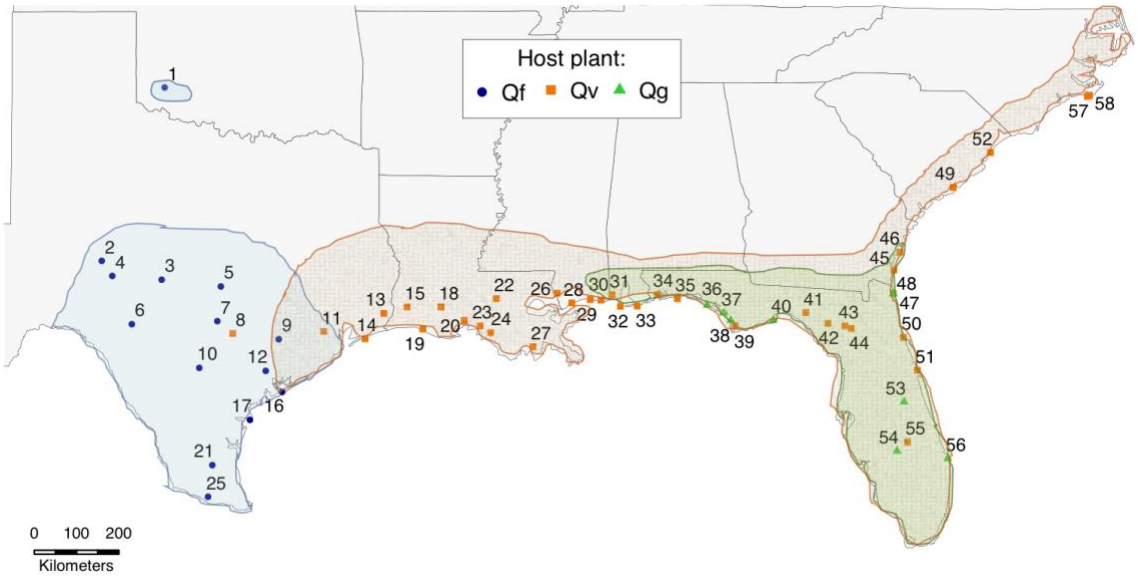

Figure 1: Sampling localities of *Belonocnema treatae* in southeastern USA. Sampling localities of *B. treatae* spanning the geographical ranges of *Q. fusiformis* (*Qf*, blue), *Q. virginiana* (*Qv*, orange) and *Q. geminata* (*Qg*, green). Numbers correspond to localities in Table 1.

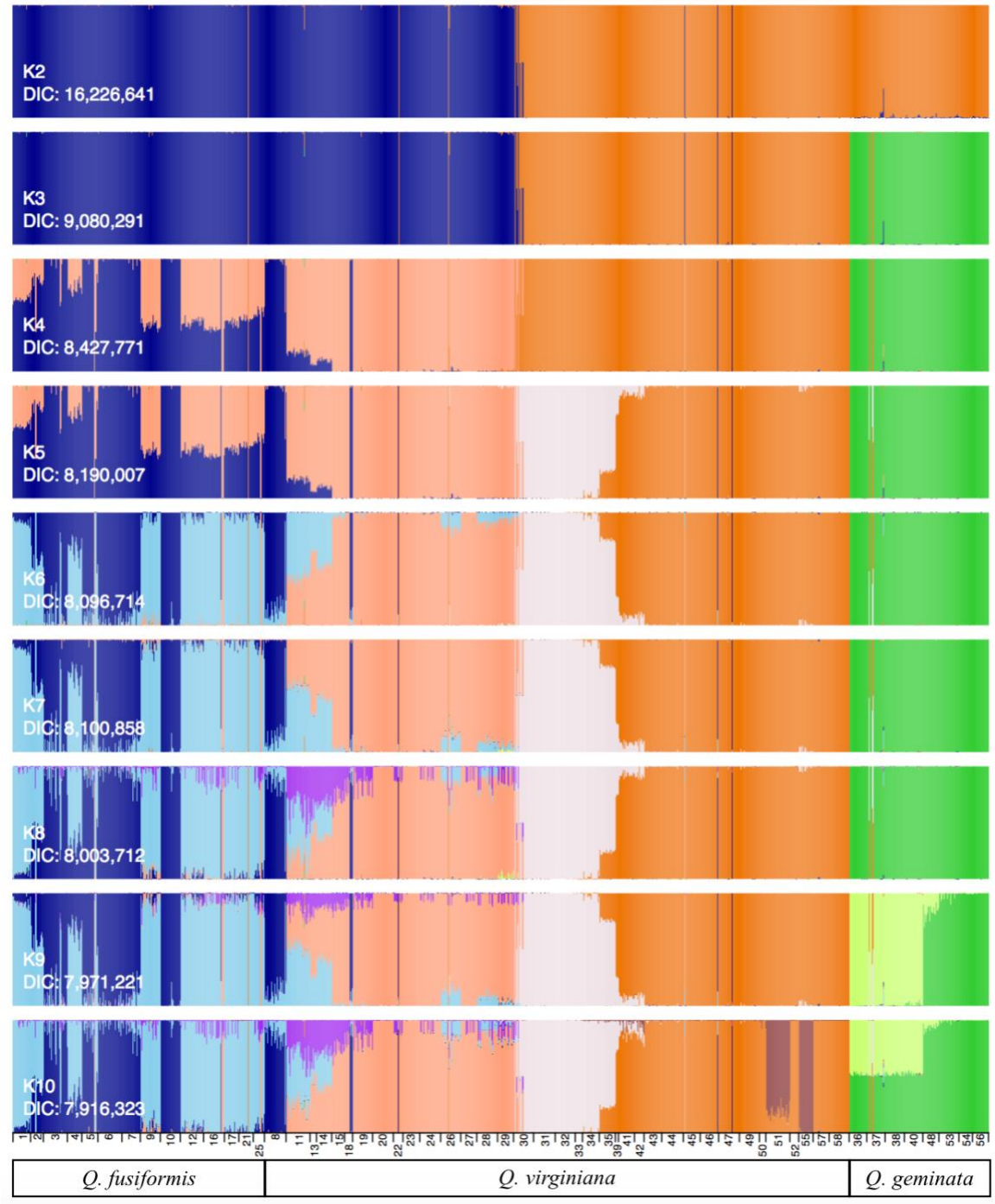

Figure 2: Admixture proportions (q) based on 1,219 individuals *k*2 – *k*10. Maximum likelihood estimates of the admixture proportion (y-axis) using 6,987 loci. Individuals are ordered by host plant affiliation and the absolute distance from the northwestern most and isolated locality (site 1, Quartz Mt., OK). The proportion of each individual's ancestry is denoted by the height of each block of color (genetic cluster). Each barplot includes the model number and DIC score.

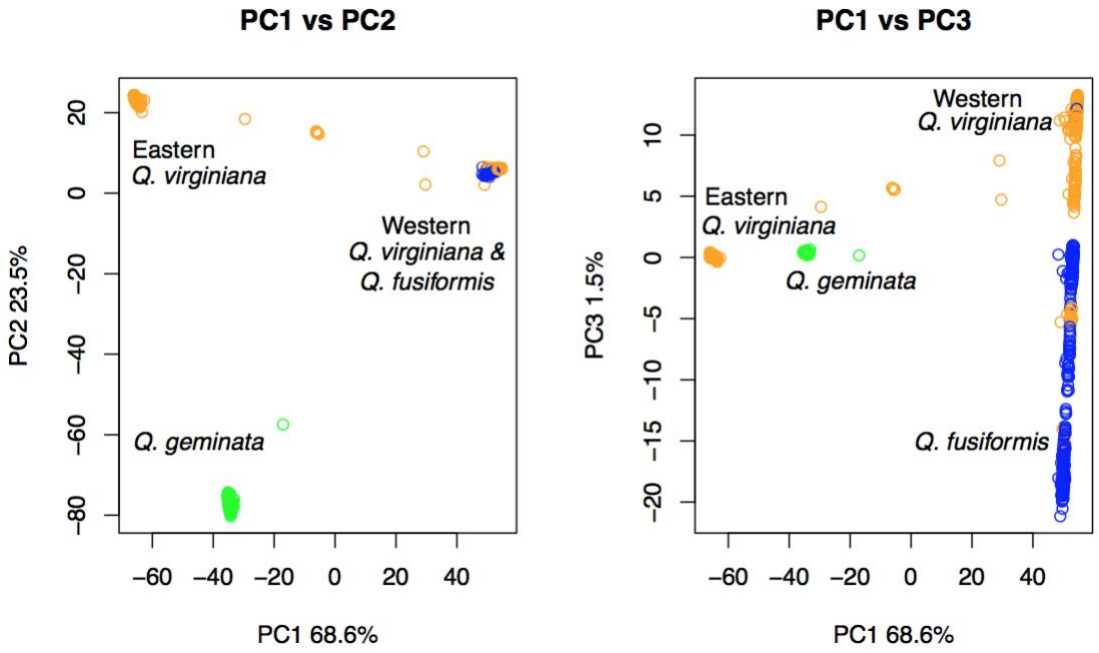

Figure 3: Genotype probability PCA. PCA plot illustrate the summaries of genomic variation (average genotype probabilities for *k*2 – *k*10) for the 6,987 loci for 1,219 *B. treatae*. Points denote individuals which and color coded by host plant affiliation: *Q. fusiformis* (blue), *Q. virginiana* (orange) and *Q. geminata* (green).

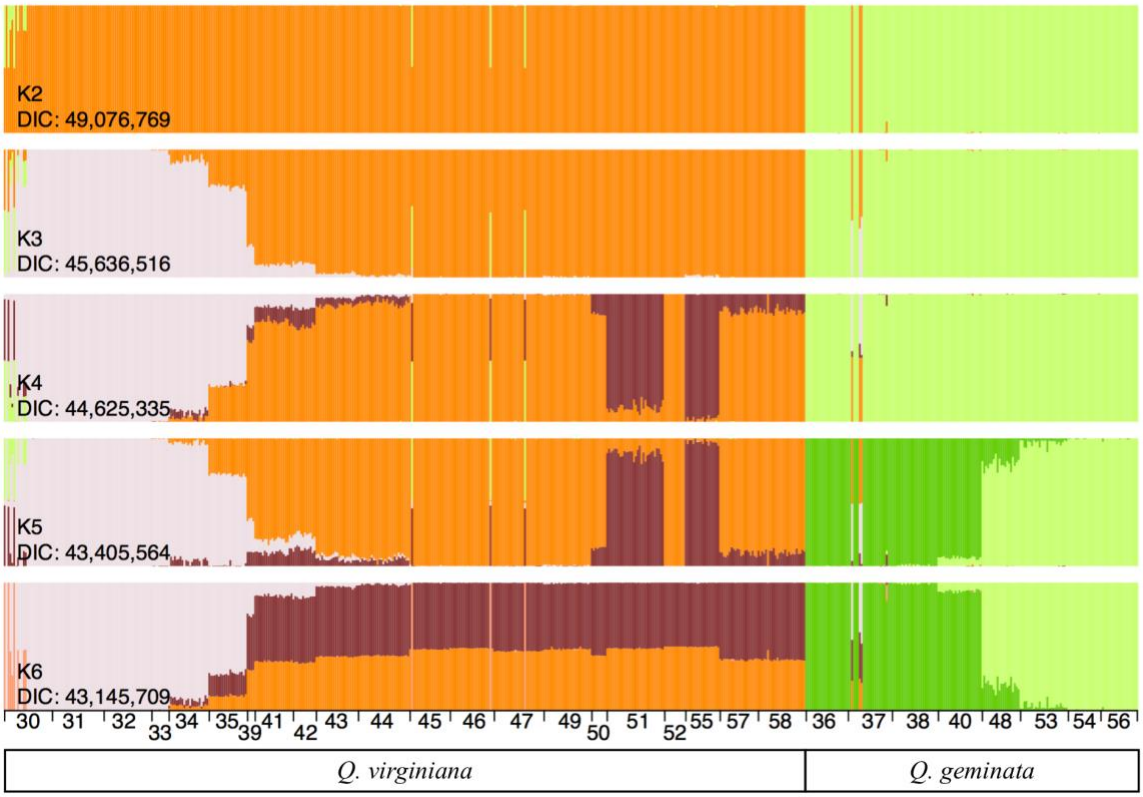

Figure 4: Admixture proportions (q) of Eastern subset *k*2 – *k*6. Proportion of ancestry estimates for the eastern subset of 593 individuals based on 38,210 loci.

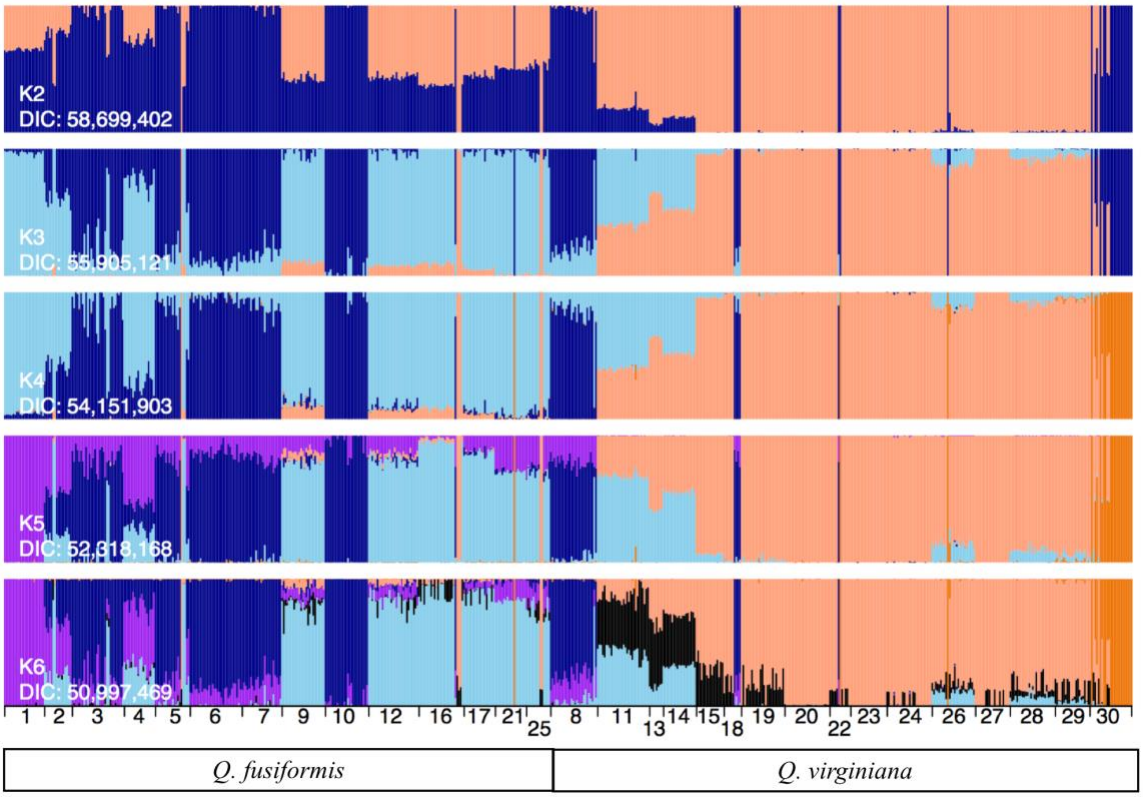

Figure 5: Admixture proportions (q) of Western subset *k*2 – *k*6. Proportion of ancestry estimates for the western subset of 651 individuals based on 25,866 loci.

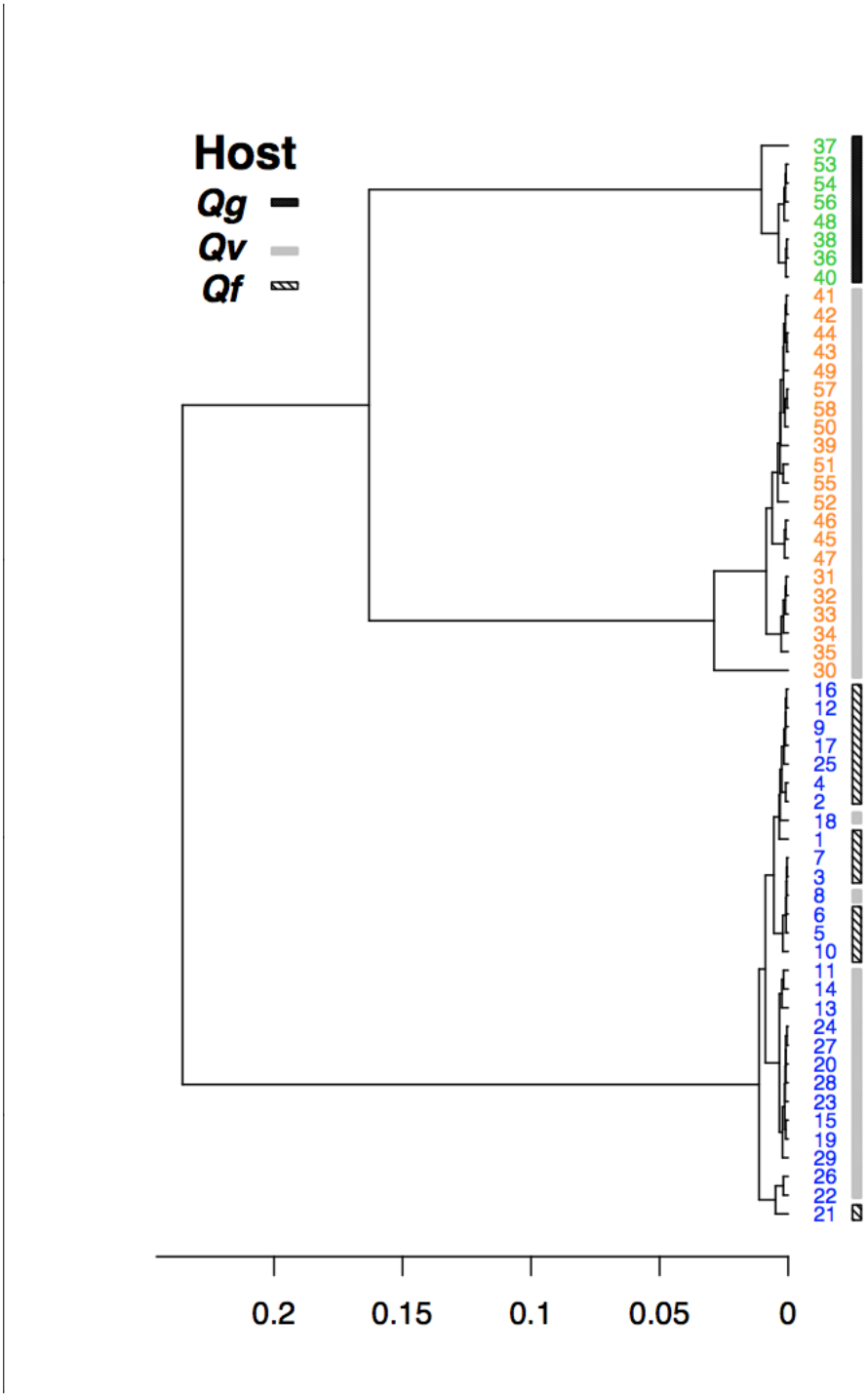

Figure 6: Nei's D_A genetic distances. Estimates of pairwise genomic distance (Nei's D_A) compared by collection locality, based on 1219 individuals with 6,987 loci. Numbers correspond to collection localities; with host association to the right of the site name.

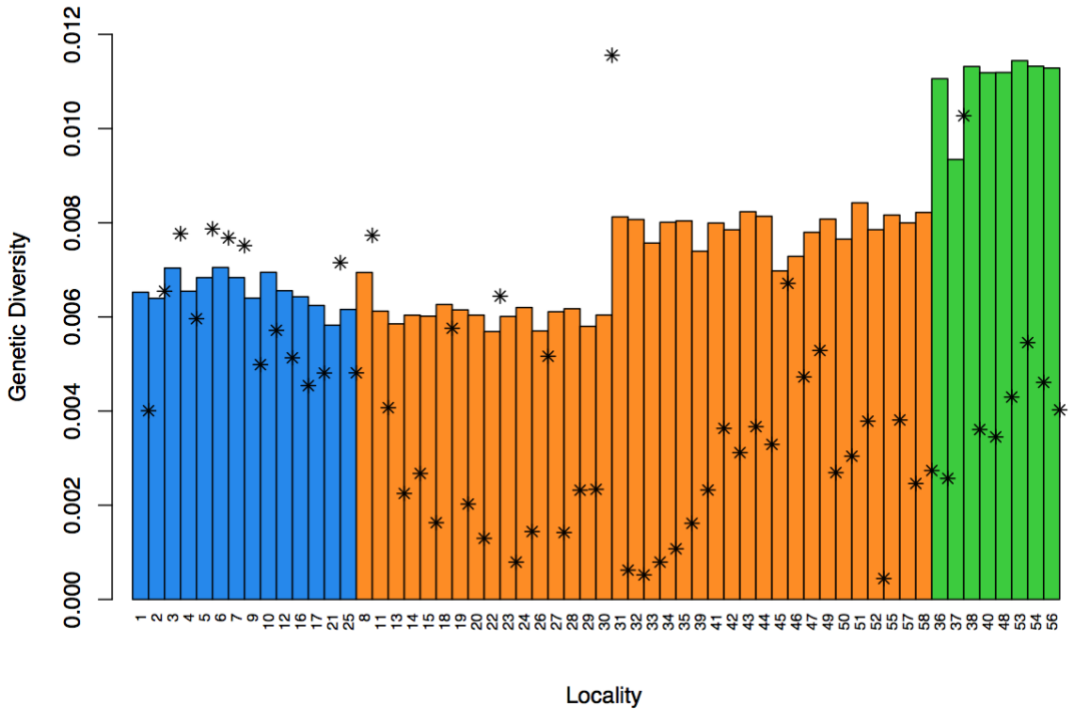

Figure 7: **Estimates of genomic diversity: Watterson's θ and Tajima's π.** Estimates of Watterson's θ are denoted by bars; estimates of expected heterozygosity (π) by $*$. Populations are in the same order as *entropy* barplots; colors correspond to host affiliations: blue (*Q. fusiformis*), orange (*Q. virginiana*) and green (*Q. geminata*).

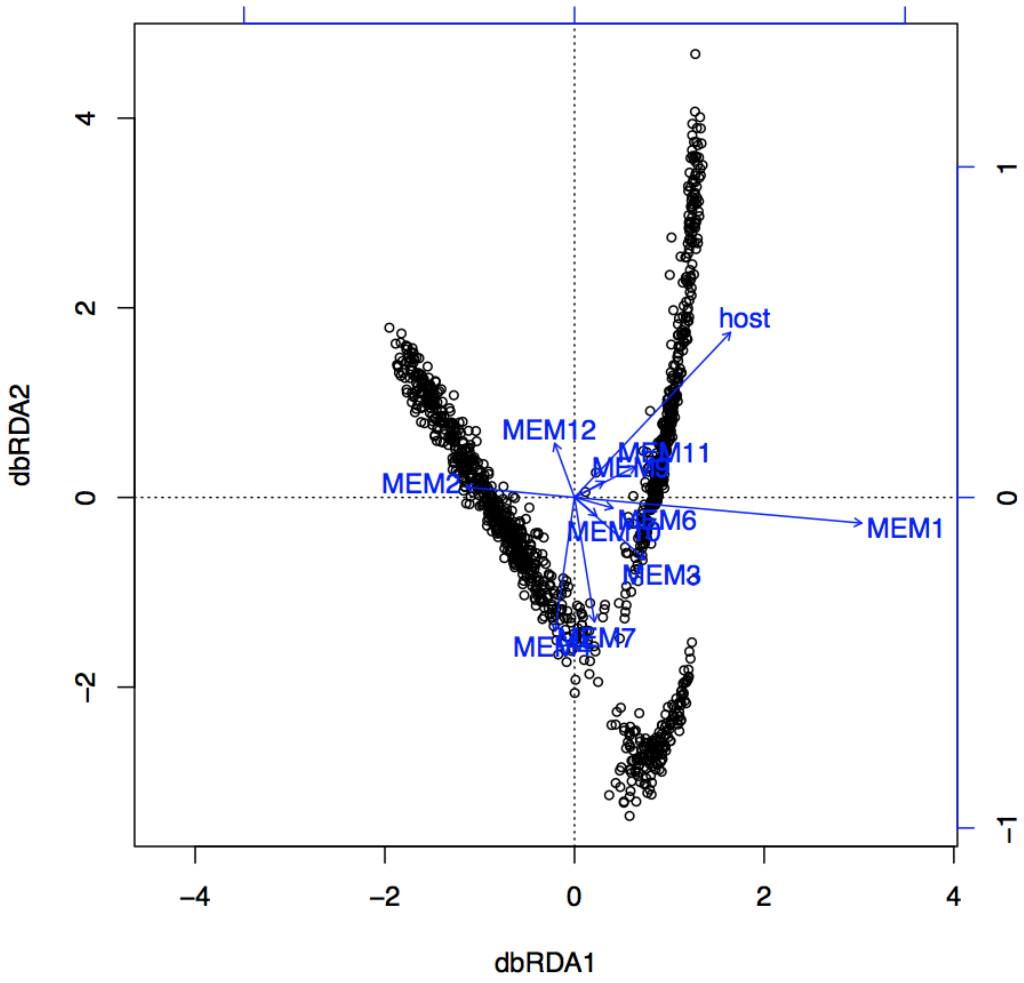

Figure 8: dbRDA: RDA1 vs RDA2. dbRDA analysis on collapsed genotype probability matrix using host affiliation and 11 significant positive MEM spatial variables as predictors. Dots indicate individuals and the length of the arrows reflect the magnitude of predictor loadings and the direction of the arrow represents which axes it is loading heaviest on.

REFERENCES

Askew, R. R. (1984). Biology of gall wasps. *Biology of gall insects/editor TN Ananthakrishnan*.

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

Borcard, D., & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, *153*(1-2), 51-68.

Borcard, D., Legendre, P., Avois-Jacquet, C., & Tuomisto, H. (2004). Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, *85*(7), 1826-1832.

Bush, G. L. (1969). Sympatric host race formation and speciation in frugivorous flies of the genus Rhagoletis (Diptera, Tephritidae). *Evolution*, *23*(2), 237-251.

Cavender-Bares, J., Kitajima, K., & Bazzaz, F. A. (2004). Multiple trait associations in relation to habitat differentiation among 17 Floridian oak species. *Ecological Monographs*, *74*(4), 635-662.

Cavender-Bares, J., & Holbrook, N. M. (2001). Hydraulic properties and freezing-induced cavitation in sympatric evergreen and deciduous oaks with contrasting habitats. *Plant, Cell & Environment*, *24*(12), 1243-1256.

Cavender-Bares, J. (2007). Chilling and freezing stress in live oaks (*Quercus* section *Virentes*): intra-and inter-specific variation in PS II sensitivity corresponds to latitude of origin. *Photosynthesis research*, *94*(2-3), 437-453.

Cavender-Bares, J., & Pahlich, A. (2009). Molecular, morphological, and ecological niche differentiation of sympatric sister oak species, *Quercus virginiana* and *Q. geminata* (Fagaceae). *American Journal of Botany*, *96*(9), 1690-1702.

Cavender-Bares, J., Gonzalez-Rodriguez, A., Pahlich, A., Koehler, K., & Deacon, N. (2011). Phylogeography and climatic niche evolution in live oaks (*Quercus s*eries *Virentes*) from the tropics to the temperate zone. *Journal of Biogeography*, *38*(5), 962-981.

Cavender-Bares, J., González-Rodríguez, A., Eaton, D. A., Hipp, A. A., Beulke, A., & Manos, P. S. (2015). Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): A genomic and population genetics approach. *Molecular ecology*, *24*(14), 3668-3687.

Courtney, S. P., & Forsberg, J. (1988). Host use by two pierid butterflies varies with host density. *Functional Ecology* 2(1), 67-75.

Coyne, J. A., & Orr, H. A. (2004). Speciation, vol. 37. *Sunderland, MA: Sinauer Associates*.

Darwin, C. (1859). On the origins of species by means of natural selection. *London: Murray*.

Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *ecological modelling*, *196*(3-4), 483-493.

Egan, S. P., & Ott, J. R. (2007). Host plant quality and local adaptation determine the distribution of a gall-forming herbivore. *Ecology*, *88*(11), 2868-2879.

Egan, S. P., Hood, G. R., & Ott, J. R. (2011). Natural Selection on Gall Size: Variable Contributions of Individual Host Plants to Population-Wide Patterns. *Evolution, 65*(12), 3543-3557.

Egan, S. P., Hood, G. R., Feder, J. L., & Ott, J. R. (2012a). Divergent host-plant use promotes reproductive isolation among cynipid gall wasp populations. *Biology letters, 8*(4), 605-608.

Egan, S. P., Hood, G. R., & Ott, J. R. (2012b). Testing the role of habitat isolation among ecologically divergent gall wasp populations. *International Journal of Ecology,* 2012(809897), 1-8

Egan, S. P., Hood, G. R., DeVela, G., & Ott, J. R. (2013). Parallel patterns of morphological and behavioral variation among host-associated populations of two gall wasp species. *PloS one, 8*(1), e54690.

Ehrlich, P. R., & Raven, P. H. (1969). Differentiation of populations. *Science*, 165(3899), 1228-1232.

Funk, D. J., Filchak, K. E., & Feder, J. L. (2002). Herbivorous insects: model systems for the comparative study of speciation ecology. *Genetica*, *116*(2-3), 251-267.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457-472.

Gompert, Z., & Buerkle, A. C. (2010). INTROGRESS: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, *10*(2), 378-384.

Gompert, Z., Lucas, L. K., Fordyce, J. A., Forister, M. L., & Nice, C. C. (2010). Secondary contact between Lycaeidesidas and L. melissa in the Rocky Mountains: extensive admixture and a patchy hybrid zone. *Molecular ecology*,*19*(15), 3171-3192.

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular ecology*, *23*(18), 4555-4573.

Gugger, P. F., & Cavender-Bares, J. (2013). Molecular and morphological support for a Florida origin of the Cuban oak. *Journal of Biogeography*, *40*(4), 632-645.

Hafner, M. S., & Nadler, S. A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, *332*(6161), 258-59

Hood, G. R., & Ott, J. R. (2010). Developmental plasticity and reduced susceptibility to natural enemies following host plant defoliation in a specialized herbivore. *Oecologia*, *162*(3), 673-683.

Hood, G. R., & Ott, J. R. (2017). Independent life history evolution between generations of bivoltine species: a case study of cyclical parthenogenesis. *Oecologia*, *183*(4), 1053-1064.

Hunter, M. D., & Price, P. W. (1992). Playing chutes and ladders: heterogeneity and the relative roles of bottom-up and top-down forces in natural communities. *Ecology*, *73*(3), 724-732.

Jaenike, J. (1990). Host specialization in phytophagous insects. *Annual Review of Ecology and Systematics* 21(1), 243-273.

Kuussaari, M., Singer, M., & Hanski, I. (2000). Local specialization and landscape-level influence on host use in an herbivorous insect. *Ecology*, *81*(8), 2177-2187.

Lee, C. R., & Mitchell-Olds, T. (2011). Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Molecular ecology*, 20(22), 4631-4642.

Legendre, P., & Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, *80*(2), 107-138.

Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular ecology resources*, *10*(5), 831-844.

Legendre, P., Fortin, M. J., & Borcard, D. (2015). Should the Mantel test be used in spatial analysis?. *Methods in Ecology and Evolution*, *6*(11), 1239-1247.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987-2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

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

Lund, J. N., Ott, J. R., & Lyon, R. J. (1998). Heterogony in *Belonocnema treatae* Mayr (Hymenoptera: Cynipidae). *Proceedings-Entomological Society of Washington*, *100*, 755-763.

Mandeville, E. G., Parchman, T. L., McDonald, D. B., & Buerkle, C. A. (2015). Highly variable reproductive isolation among pairs of *Catostomus* species. *Molecular ecology*, *24*(8), 1856-1872.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209-220.

Manos, P. S., Doyle, J. J., & Nixon, K. C. (1999). Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Molecular phylogenetics and evolution*, *12*(3), 333-349.

Malyshev, S. I. (1968). Genesis of the Hymenoptera. In *Genesis of the Hymenoptera and the phases of their evolution* (pp. 3-9). Springer, Boston, MA.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Miller, D. G., Ivey, C. T., & Shedd, J. D. (2009). Support for the microenvironment hypothesis for adaptive value of gall induction in the California gall wasp, *Andricus quercuscalifornicus*. *Entomologia Experimentalis et Applicata*, *132*(2), 126-133.

Mitter, C., Farrell, B., & Wiegmann, B. (1988). The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification?. *The American Naturalist*, *132*(1), 107-128.

Muller, C. H. (1961). The live oaks of the series *Virentes.* American Midland Naturalist, 65(1), 17-39.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, *70*(12), 3321-3323.

Nei, M., & Takezaki, N. (1983). Estimation of genetic distances and phylogenetic trees from DNA analysis. *Proc 5th World Cong Genet Appl Livstock Prod*, *21*(21), 405-412.

Nixon K.C. (1985) A Biosystematic Study of *Quercus* Series *Virentes* (the live oaks) with Phylogenetic Analyses of Fagales, Fagaceae and Quercus, Ph.D. Thesis. University of Texas, Austin.

Nychka, D., & Nychka, M. D. (2017). Package 'LatticeKrig'.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular ecology*, *21*(12), 2991-3005.

Parchman, T. L., Buerkle, C. A., Soria-Carrasco, V., & Benkman, C. W. (2016). Genome divergence and diversification within a geographic mosaic of coevolution. *Molecular ecology*, *25*(22), 5705-5718.

Pearse, I. S., & Hipp, A. L. (2009). Phylogenetic and trait similarity to a native species predict herbivory on non-native oaks. *Proceedings of the National Academy of Sciences*, *106*(43), 18097-18102.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news*, *6*(1), 7-11.

Price PW, Fernandes GW, Waring GL. 1987. Adaptive nature of insect galls. *Environ. Entomol.* 16(1),15–24

Quicke, D. L. (1997). *Parasitic wasps*. Chapman & Hall Ltd.

Radersma, R., Garroway, C. J., Santure, A. W., De Cauwer, I., Farine, D. R., Slate, J., & Sheldon, B. C. (2017). Social and spatial effects on genetic variation between foraging flocks in a wild bird population. *Molecular ecology*.

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology letters*, *8*(3), 336-352.

Schluter, D. (1996). Adaptive radiation along genetic lines of least resistance. *Evolution*, 1766-1774.

Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution*, *16*(7), 372-380.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, *323*(5915), 737-741.

Schuler, H., Egan, S. P., Hood, G. R., Busbee, R. W., Driscoe, A. L., & Ott, J. R. (2018). Diversity and distribution of Wolbachia in relation to geography, host plant affiliation and life cycle of a heterogonic gall wasp. *BMC evolutionary biology*, *18*(37), 1-15

Servedio, M. R. (2016). Geography, assortative mating, and the effects of sexual selection on speciation with gene flow. *Evolutionary applications*, *9*(1), 91-102.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693-702.

Stone, G. N., Schönrogge, K., Atkinson, R. J., Bellido, D., & Pujade-Villar, J. (2002). The population biology of oak gall wasps (Hymenoptera: Cynipidae).*Annual review of entomology*, *47*(1), 633-668.

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, *144*(1), 389-399.

Tooker, J. F., & De Moraes, C. M. (2008). Gall insects and indirect plant defenses: A case of active manipulation? *Plant signaling & behavior*, *3*(7), 503-504.
Turelli, M., & Hoffmann, A. A. (1991). Rapid spread of an inherited incompatibility factor in California Drosophila. *Nature*, *353*(6343), 440-42.
Turelli, M., Hoffmann, A. A., & McKechnie, S. W. (1992). Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics*, *132*(3), 713-723.
Underwood, N., & Rausher, M. D. (2000). The effects of host-plant genotype on herbivore population dynamics. *Ecology*, *81*(6), 1565-1576.
Wang, I. J., Glor, R. E., & Losos, J. B. (2013). Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology letters*, *16*(2), 175-182.
Werren, J. H. (1997). Biology of *Wolbachia*. *Annual review of entomology*, *42*(1), 587-609.
Winkler, I., Scheffer, S.J., Lewis, M.L., Ottens, K.J., Rasmussen, A.P., Gomes-Costa, G.A., Santillan, L.M.H., Condon, M.A. and Forbes, A.A. (2018). Anatomy of a Neotropical insect radiation. *BMC evolutionary biology*, *18*(1), 1-13.
Wright, S. (1942). Statistical genetics and evolution. *Bulletin of the American Mathematical Society*, *48*(4), 223-246.
Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114-138.
Wiklund, C. (1974). The concept of oligophagy and the natural habitats and host plants of *Papilio machaon* L. in Fennoscandia. *Insect Systematics & Evolution*, *5*(2), 151-160.
Zhang, L., Driscoe, A., Izen, R., Toussaint, C., Ott, J. R., & Egan, S. P. (2017). Immigrant inviability promotes reproductive isolation among host-associated populations of the gall wasp *Belonocnema treatae*. *Entomologia Experimentalis et Applicata*, *162*(3), 379-388.