Fall 2018

Population Structuring and Phylogenetic Inference in a Species of Tropical Fly Using Double Digest Rad Markers

Dacia Lipkea

Follow this and additional works at: https://ir.uiowa.edu/honors_theses
Part of the Ecology and Evolutionary Biology Commons

Copyright © 2018 Dacia Lipkea

Hosted by Iowa Research Online. For more information please contact: lib-ir@uiowa.edu.
POPULATION STRUCTURING AND PHYLOGENETIC INFERENCE IN A SPECIES OF TROPICAL FLY USING DOUBLE DIGEST RAD MARKERS

by

Dacia Lipkea

A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Biology

Andrew Forbes; Robin Bagley
Thesis Mentor

Fall 2018

All requirements for graduation with Honors in the Biology have been completed.

Lori Adams
Biology Honors Advisor

This honors thesis is available at Iowa Research Online: https://ir.uiowa.edu/honors_theses/
POPULATION STRUCTURING AND PHYLOGENETIC INFERENCE IN A SPECIES OF TROPICAL FLY USING DOUBLE DIGEST RAD MARKERS

by

Dacia Lipkea

A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Department of Biology

Andrew Forbes, PhD
Thesis Mentor

Fall 2018

All requirements for graduation with Honors in the Department of Biology have been completed.

Lori Adams, PhD.
Biology Honors Advisor
Abstract

*Blepharoneura* (Diptera: Tephritidae) is a highly diverse, Neotropical genus of fruit flies that feed on cucurbit (Cucurbitaceae) host plants. Like most plant-feeding insects, *Blepharoneura* are highly specialized, with most species utilizing only a single part of one host species. Due to this extreme specialization and close, life-long associations with their host plants, shifts and subsequent adaptation to new hosts would be expected to drive divergence and diversification within the genus. However, sister species of *Blepharoneura* occupy extremely overlapping niches – they frequently share host plants and often utilize the same plant tissues. To investigate what role, if any, host use plays in driving divergence, we use double-digest restriction-associated sequencing (ddRAD-seq) to examine population structure and phylogenetic relationships between individuals of *Blepharoneura* species 10, which uses multiple host plants and parts throughout its range. Our findings show that, although the majority of divergence is linked to geographic location, host use may also contribute to differentiation.
Acknowledgements

I would like to thank all members of the Forbes lab for their continual support, advice, and assistance during the execution of this project. A special thanks to Heather Widmayer and Robin Bagley for being so awesome and helping me through seemingly endless troubleshooting! I would also like to thank past and present collaborators on the Blepharoneura project who have participated in specimen collection, curation, DNA extractions, discussions, etc. This work was supported by a University of Iowa ICRU 2018 Summer Research Fellowship and by NSF DEB1542269.
# Table of Contents

List of Tables ........................................................................................................................................... IV

List of Figures ............................................................................................................................................... V

1. Introduction .............................................................................................................................................. 1

2. Materials and methods ............................................................................................................................ 5
   2.1. Sample collections ............................................................................................................................. 5
   2.2. Double-digest restriction-associated DNA library preparation and sequencing ................................ 5
   2.3. Processing raw sequence reads ......................................................................................................... 8
   2.4. Inference of population structure ..................................................................................................... 9
   2.5. SNP-based phylogenetic analyses ..................................................................................................... 10
   2.6. Locus-based phylogenetic analyses ................................................................................................. 12

3. Results ...................................................................................................................................................... 14
   3.1. Illumina sequencing data .................................................................................................................... 14
   3.2. Inference of population structure ..................................................................................................... 14
   3.3. Phylogenetic analyses ....................................................................................................................... 18

4. Discussion ................................................................................................................................................ 27

5. Literature Cited ....................................................................................................................................... 34
List of Tables

Table 1. Sampling data for the 33 Blepharoneura sp. 10 individuals used in this study. ............7

Table 2. Number of RAD loci and SNPs missing from 30% or 50% of all 33 Blepharoeura sp. 10 individuals .................................................................15
List of Figures

Figure 1. *Blepharoneura* sp. 10 sampling. ................................................................. 6

Figure 2. C.V. error plots.................................................................................................. 16

Figure 3. Population structure under a model of $K = 2$. ................................................ 17

Figure 4. Phylogenetic network inferred with RAxML, based on concatenated RAD SNPs for the 50% dataset. ......................................................................................... 19

Figure 5. Phylogenetic network inferred with RAxML, based on concatenated RAD SNPs for the 30% dataset. ......................................................................................... 20

Figure 6. Phylogenetic network inferred with quartet-based analysis, as implemented in PAUP*, based on concatenated RAD SNPs for the 50% dataset. .................................................. 21

Figure 7. Phylogenetic network inferred with quartet-based analysis, as implemented in PAUP*, based on concatenated RAD SNPs for the 30% dataset. .................................................. 22

Figure 8. Phylogenetic network inferred with RAxML, based on concatenated RAD loci for the 50% dataset. ......................................................................................... 23

Figure 9. Phylogenetic network inferred by RAxML, based on concatenated RAD loci for the 30% dataset. ......................................................................................... 24

Figure 10. Phylogenetic network inferred using Bayesian inference, based on concatenated RAD loci for the 50% dataset. ......................................................................................... 25

Figure 11. Phylogenetic network inferred using Bayesian inference, based on concatenated RAD loci for the 30% dataset. ......................................................................................... 26
1. Introduction

A large proportion of species on Earth today are phytophagous insects that specialize on a specific part of one or a handful of host plant species (Strong et al, 1984; May, 1990; Wilson, 1992). Comparative studies of sister taxa find plant-feeding lineages are more diverse than those with other lifestyles (Mitter, Farrell, & Wiegmann, 1988; Farrell, 1998); but the reason for this increased diversity has been long debated. Although several hypotheses have been offered (e.g., increased speciation rate, decreased extinction rate), the link between plant-feeding and species diversity is somewhat unclear (Janz et al, 2006).

One hypothesis is that shifts to new host species promote speciation (Jaenike, 1990; Matsubayashi, 2010). This is thought to be true because insects spend their lives intimately associated with their host: they rely on their host for a mating location, egg laying location, and a food source for larvae. Many also pupate on or near their host as they develop into adults and continue the cycle. Since the majority of plant-feeding insects are not only intimately associated with, but also are highly specialized, shifts and subsequent adaptations to new hosts can lead to the accumulation of reproductive isolating barriers as a by-product and ultimately result in speciation (Nosil, 2002; Matsubayashi, 2010; Forbes et al, 2017).

A method by which this occurs is through selection on existing or novel variation. All individuals in a population contain some degree of genetic variation, which can be increased by mutation and outcrossing events with other populations. If members of this population migrate to a new environment or habitat, natural selection can act on their genetic variation and result in fine-tuned adaptations to the host plant they colonize in the new environment. As an insect continues to use the resources provided by a new host, these adaptations will subsequently lead to the slow accumulation of genetic differences from the original population. If this results in a
higher fitness, individuals may specialize on the new host plant and avoid individuals that occupy the parental host plant, diverging even further from the parental taxa (Fry, 1999). Host fidelity and subsequent assortative mating may then drive disruptive selection, and when disruptive selection is stronger than gene flow between populations that lack extrinsic barriers to gene flow, reproductive isolation is more likely to occur (Sezer & Butlin, 1998). Reproductive isolation, then, in the form of either pre- or postzygotic isolation, results in a further increase in the number of genetic differences until the locally adapted subpopulation can no longer create viable offspring with the parental species (Orr & Smith, 1998).

In addition to specializing on a particular host species, insects often utilize a specific part of their host plant. A single plant contains numerous diverse habitats, such as leaves, stems, flowers, fruits, and seeds, each of which can be independently colonized by different species. Additionally, the parts of a plant are generally spatially isolated and can also be temporally isolated, creating even more opportunities for different colonizations. For dioecious plants, in particular, it is common to see male and female inflorescences that flower at different times during the year (Condon et al, 2008). Therefore, environments containing numerous different plant species offer a wide variety of specific niches that can be utilized by specialized, plant-feeding insects. The coexistence of many phytophagous species with niches in close proximity provides many more opportunities for host shifts to occur.

An example of a biological system in which we might anticipate divergence to occur following host shifts is *Blepharoneura*. *Blepharoneura* is a Neotropical genus of Tephritid fruit flies that specialize on cucurbit host plants (Cucurbitaceae). Most cucurbit species are monoecious and have male and female reproductive organs located on separate, sexually dimorphic branches (Condon & Gilbert, 1988, 1990). Most *Blepharoneura* species specialize on
the calyx tissue of flowers of one sex of one host cucurbit species, although some species associate with seeds instead of flowers. Female flies will lay eggs in the host tissue, which larvae feed on throughout development. Larvae typically remain in the host tissue until it falls to the ground, at which point they pupate in the soil and eventually emerge as mature adults.

Because plant-feeding insects are typically highly specialized on one part of the host plant, it is uncommon to see different species sharing the same niche, or host plant part. In *Blepharoneura*, however, we see extreme overlap of niches in which individual species utilize multiple host parts that are frequently already occupied by other *Blepharoneura* species. For example, while studying the diversity of individuals at a collection site in Ecuador, six different *Blepharoneura* species, identified by their cytochrome oxidase I mitochondrial haplotype, were found to occupy the same sex flower of the same host species (Condon et al, 2008). The sharing of the same host plant tissue by multiple fly species calls into question whether host shifts play a role in driving divergence in *Blepharoneura*.

Determining drivers of divergence can be difficult when using between-species comparisons, as post-speciational changes can accrue and obscure initial patterns of divergence (Coyne and Orr 2004). Instead, it may be more useful to search within species found on multiple different host plant species and/or parts for evidence of incipient divergence. By collecting many individuals of a single species from their full geographic distribution and from as much of their host variation as possible, we may begin to understand what is allowing this genus of flies occupy extremely overlapping niches and detect signals of divergence.

To that end, here we examine genetic differentiation within *Blepharoneura* species 10 (hereafter, “sp. 10”). Sp. 10 is distributed throughout northern South America, and although primarily associated with female flowers of *Gurania spinulosa*, has also been collected on
several other plant species, and on both male and female flowers. This diversity in host affiliation is unusual within *Blepharoneura* species, and so may help us to uncover the role of host use in diversification unclear for this system. Using markers generated with a double-digest restriction-associated sequencing (ddRAD-seq) approach, we look for evidence of divergence within sp. 10 using both population structuring and phylogenetic analyses. We also investigate the robustness of our results by repeating our analyses using multiple filtering strategies and computational approaches.
2. Materials and methods

2.1. Sample collections

Specimens of Blepharoneura were collected over 15 years from multiple locations across South America (Figure 1). Flowers and fruits were collected from cucurbit host plants (Cucurbitaceae). For plants that were out of arm’s reach, a 12m collecting pole was used to collect specimens. If there were plants beyond the reach of the collecting pole, the ground was searched for fallen flowers or fruit. All flowers and fruits found at a site were collected and the coordinates of each collection site were recorded. Each flower and fruit that was collected was placed singly into plastic cups that were checked each day to determine if any larvae had emerged. Whenever larvae emerged, they were placed individually into another set of plastic cups containing a moist substrate until they emerged as adults. The adult Blepharoneura were fed sugar water until they were fully developed, at which point they were killed and stored in 95% ethanol at -80°C (see Condon et al, 2008).

2.2. Double-digest restriction-associated DNA library preparation and sequencing

DNA was extracted from 33 preserved adults putatively identified as Blepharoneura sp. 10 based on mitochondrial COI haplotype (Figure 1; Table 1; Condon et al, 2008, 2014). Genomic DNA was obtained using either a Qiagen DNeasy kit (Qiagen Inc., Valencia, CA, USA) or a CTAB extraction method modified from Chen et al. (2010). The DNA was then quantified using a Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen – Molecular Probes, Eugene, OR, USA).

To generate a large set of genome-wide markers for analysis, we took a double digest restriction-associated DNA (ddRAD) sequencing approach (Peterson et al. 2012). Following the protocol outlined in Peterson et al. 2012, we performed a series of test restriction digests to select
Figure 1. *Blepharoneura* sp. 10 sampling. Location and host flower affiliation for each of the 33 specimens in this study. Individuals were selected to represent the full geographic distribution and as much of the host variation as possible.
Table 1. Sampling data for the 33 *Blepharoneura* sp. 10 individuals used in this study.

<table>
<thead>
<tr>
<th>Collection ID</th>
<th>Working ID</th>
<th>Host Plant</th>
<th>Flower Sex</th>
<th>Country</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blar204</td>
<td>P1</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.52</td>
<td>-73.09</td>
</tr>
<tr>
<td>Blar221</td>
<td>P2</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.52</td>
<td>-73.09</td>
</tr>
<tr>
<td>Blar26</td>
<td>T1</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Trinidad</td>
<td>10.59</td>
<td>-61.25</td>
</tr>
<tr>
<td>Blar278</td>
<td>P3</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.53</td>
<td>-73.18</td>
</tr>
<tr>
<td>Blar282</td>
<td>P4</td>
<td><em>G. acuminata</em></td>
<td>Female</td>
<td>Peru</td>
<td>-4.97</td>
<td>-73.65</td>
</tr>
<tr>
<td>Blar283</td>
<td>P5</td>
<td><em>G. acuminata</em></td>
<td>Female</td>
<td>Peru</td>
<td>-4.97</td>
<td>-73.65</td>
</tr>
<tr>
<td>Blar284</td>
<td>P6</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.53</td>
<td>-73.18</td>
</tr>
<tr>
<td>Blep1202</td>
<td>S1</td>
<td><em>G. robusta</em></td>
<td>Female</td>
<td>Suriname</td>
<td>4.95</td>
<td>-55.18</td>
</tr>
<tr>
<td>Blep14</td>
<td>V1</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Venezuela</td>
<td>10.07</td>
<td>-66.59</td>
</tr>
<tr>
<td>Blep214</td>
<td>E1</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Ecuador</td>
<td>0.38</td>
<td>-79.78</td>
</tr>
<tr>
<td>Blep303</td>
<td>F1</td>
<td><em>G. bignoniacea</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.25</td>
<td>-52.13</td>
</tr>
<tr>
<td>Bpup1205</td>
<td>F2</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.15</td>
<td>-52.14</td>
</tr>
<tr>
<td>Bpup1316</td>
<td>F3</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.15</td>
<td>-52.10</td>
</tr>
<tr>
<td>Bpup1367</td>
<td>F4</td>
<td><em>G. spinulosa</em></td>
<td>Male</td>
<td>French Guiana</td>
<td>4.70</td>
<td>-52.31</td>
</tr>
<tr>
<td>Bpup1373</td>
<td>F5</td>
<td><em>G. spinulosa</em></td>
<td>Male</td>
<td>French Guiana</td>
<td>4.70</td>
<td>-52.31</td>
</tr>
<tr>
<td>Bpup1625</td>
<td>F6</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.16</td>
<td>-52.15</td>
</tr>
<tr>
<td>Bpup1834</td>
<td>F7</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.60</td>
<td>-52.27</td>
</tr>
<tr>
<td>Bpup1888</td>
<td>F8</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.57</td>
<td>-52.16</td>
</tr>
<tr>
<td>Bpup1906</td>
<td>F9</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.15</td>
<td>-52.14</td>
</tr>
<tr>
<td>Bpup1915</td>
<td>F10</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>French Guiana</td>
<td>4.15</td>
<td>-52.14</td>
</tr>
<tr>
<td>Bpup2172</td>
<td>S2</td>
<td><em>G. robusta</em></td>
<td>Unknown</td>
<td>Suriname</td>
<td>4.95</td>
<td>-55.18</td>
</tr>
<tr>
<td>Bpup2584</td>
<td>P7</td>
<td><em>G. acuminata</em></td>
<td>Female</td>
<td>Peru</td>
<td>-4.50</td>
<td>-73.59</td>
</tr>
<tr>
<td>Bpup2800</td>
<td>P8</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.53</td>
<td>-73.18</td>
</tr>
<tr>
<td>Bpup2928</td>
<td>P9</td>
<td><em>G. acuminata</em></td>
<td>Female</td>
<td>Peru</td>
<td>-4.49</td>
<td>-73.60</td>
</tr>
<tr>
<td>Bpup3178</td>
<td>P10</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-3.53</td>
<td>-73.18</td>
</tr>
<tr>
<td>Bpup435</td>
<td>P11</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-12.56</td>
<td>-70.11</td>
</tr>
<tr>
<td>Bpup454</td>
<td>P12</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-12.56</td>
<td>-70.11</td>
</tr>
<tr>
<td>Bpup62</td>
<td>P13</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-12.56</td>
<td>-70.11</td>
</tr>
<tr>
<td>Bpup78</td>
<td>P14</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Peru</td>
<td>-12.56</td>
<td>-70.11</td>
</tr>
<tr>
<td>KOF145</td>
<td>B1</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Bolivia</td>
<td>-16.36</td>
<td>-62.00</td>
</tr>
<tr>
<td>KOF209</td>
<td>P15</td>
<td><em>G. spinulosa</em></td>
<td>Male</td>
<td>Peru</td>
<td>-12.56</td>
<td>-70.11</td>
</tr>
<tr>
<td>MC20</td>
<td>E2</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Ecuador</td>
<td>-1.06</td>
<td>-77.60</td>
</tr>
<tr>
<td>MC52</td>
<td>E3</td>
<td><em>G. spinulosa</em></td>
<td>Female</td>
<td>Ecuador</td>
<td>-1.06</td>
<td>-77.60</td>
</tr>
</tbody>
</table>
an enzyme pair for use in this system. Based on *Blepharoneura’s* estimated genome size from flow cytometry (~300-360MB), and the results of the test digests, we selected the enzymes NlaIII (…CATGˇ…; NEB, Ipswich, MA, USA) and EcoRI (…GˇAATTC…; NEB). Following the protocol outlined in Bagley et al. (2017), we digested 200ng of DNA from each sample using these enzymes, and then ligated a universal, biotin-labeled “P2” adapter and one of 48 unique barcoded “P1” adapters to the sticky ends of each specimen. Following barcoding, we pooled the samples, and performed size selection of a 378bp fragment (± 38 bp) using a BluePippin automated gel excision system (Sage Science, Beverly, MA, USA). Following size selection, we removed fragments lacking the P2 adapter using a streptavidin-labeled Dynabeads (Life Technologies AS, Oslo, Norway). We then amplified the cleaned library using 12-rounds of high-fidelity PCR (Phusion High-Fidelity DNA polymerase, NEB) using primers that included an 8-bp Illumina multiplexing index and four degenerate bases for PCR duplicate detection. After verifying successful amplification on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA), the library was sequenced at the University of Iowa’s Iowa Institute for Human Genetics on a HiSeq 4000 using 2x150bp paired-end reads.

2.3. Processing raw sequence reads

We processed the raw Illumina reads from the 33 individuals using STACKS’ *process_radtags* program, which demultiplexes (or parses out) reads belonging to different individuals based on their barcodes, removes adapter sequences, and filters out low-quality reads based on a quality scores (v1.46; Catchen et al. 2013). Because no reference genome is available for *Blepharoneura* or any other closely related species, we used STACKS’ *denovo_map.pl* wrapper script to process and de novo assemble the reads into loci. The *denovo_map.pl* is a wrapper script that executes the STACKS pipeline by individually running the components
ustacks, cstacks, and sstacks. The ustacks component takes sets of short-read sequences for each individual and sorts them into stacks containing exactly matching sequences to form alleles. With these putative alleles, ustacks then creates a set of putative loci to find variable sites within each locus. cstacks takes the output of ustacks to merge the putative alleles together to create a catalog of consensus loci. sstacks then matches the loci within each individual to those included in the catalog. After examining a number of different assembly parameters, our final assembly utilized a stack depth (-m) of 10, allowed up to 2 mismatches between stacks to form putative loci (-M), and allowed 1 mismatch between stacks during catalog construction (-n).

We used STACKS’ populations module to generate final datasets. To examine the impact of missing data on phylogenetic and population genetic results, we produced datasets allowing either 50% or 30% missing data (-r 0.5 or 0.7, respectively) (hereafter, 50% dataset and 30% dataset, respectively). For each of these datasets, we also generated output files containing only single-nucleotide polymorphism (SNP) data, and files containing the full sequence of the RAD loci. For the SNP datasets, we utilized the “--write_random_snp” and “--phylip_var” options to create an output file in PHYLIP format containing only one random SNP per RAD locus encoded using IUPAC notation. To generate the two datasets containing RAD loci, we first had to create a population map in which each of the 33 samples was its own “population.” We then used the “--phylip_var_all” option to output the entire sequence length of each locus (SNP and invariant positions) in PHYLIP format for downstream analysis.

2.4. Inference of population structure

To determine the underlying population genetic structure of the 33 sp. 10 samples, we ran our two SNP datasets through the maximum-likelihood clustering algorithm ADMIXTURE (v1.3.0; Alexander et al. 2009). ADMIXTURE simultaneously assesses the fit of data to a model of K
populations and determines the ancestry proportion of each individual to each of the populations. We tested values of $K$ from 1 to 10 and performed 100 independent runs for each $K$ value. We determined the optimal $K$ value for each dataset by examining the average CV error score from the 100 runs for each value of $K$ as described in the *admixture* manual. We then used the main pipeline of *clumpak* (v1.1; Kopelman et al. 2015) to summarize the solutions of the 100 runs for each value of $K$. Finally, we visualized the solutions for the optimal value of $K$, and for other values of $K$ near the optimal value using a custom script in RStudio (Rstudio Team, 2016).

**2.5. SNP-based phylogenetic analyses**

For our two SNP datasets, we inferred phylogenetic relationships between the 33 individuals using both a maximum-likelihood approach (RAxML; v8.2.11; Stamatakis, 2014) and a quartet-based approach (SVDquartets; v4.0a162; Chifman & Kubatko, 2014). To determine the optimal RAxML parameter settings for our data, we first followed along the steps outlined in Stamatakis (2015) for the 30% dataset. After generating trees with the GTRCAT and GTRGAMMA models of evolution, we determined that the likelihood scores of the best GTRGAMMA tree was larger and thus GTRGAMMA was the best fitting evolutionary model. We then inferred GTRGAMMA trees with the slow search algorithm and determined that the faster, default algorithm has the best tree with the highest likelihood score. When creating these trees, RAxML by default uses randomized stepwise addition parsimony (RAP) trees as starting points for maximum likelihood tree searches but also has the option to start with completely random trees. Thus, we conducted a maximum likelihood tree search using complete random starting trees, denoted by the data filtering option “–d”, and found that the best tree produced is more supported than the best tree produced by the RAP trees for our 30% dataset.
We followed the same steps for our 50% dataset and found that the same parameter settings were optimal, except the RAP trees produced a more optimal best tree than when using complete random starting trees. After determining the best maximum likelihood tree for the two datasets we conducted a bootstrap analysis for each. Instead of specifying the specific number of bootstrap replicates we wanted to compute, we used the “-# autoMRE” data filtering option, which internally determines an appropriate number of bootstrap replicates to use and saves on some computational time (Stamatakis, 2015). After inferring bootstrap replicate trees we drew the bootstrap support values onto the best maximum likelihood tree and visualized the resulting radial tree with FigTree (v1.4.3; available at: http://tree.bio.ed.ac.uk/software/figtree/).

As SVDquartets is implemented in PAUP* (v4.0a159; Swofford, 2003), we first converted our VCF file into a NEXUS file where all SNPs for each individual were concatenated using PGDSpider (v2.1.1.5; Lisher & Excoffier, 2012). SVDquartets is a beneficial phylogenetic inference method to use because it is a single site method that avoids the commonly inaccurate gene tree estimation for each locus, especially in datasets containing few variable sites. For each combination of four sp. 10 individuals, SVDquartets generated three quartet trees corresponding to the three possible topologies. It then assigned an SVD score to each of the three topologies, and the topology with the lowest score was the true topology for that particular quartet (Chou et al, 2015). After the best quartet tree was selected for all possible combinations of four individuals, SVDquartets created a species tree that aligned to as many of the quartet trees as possible. A Maximum Quartet Support Species Tree (MQSST) score, representing the total number of quartet trees in agreement with the species tree, was then assigned to the species tree (Vachaspati & Warnow, 2018). A heuristic search to find the species tree with the highest
MQSST score was then implemented with PAUP* and the resulting radial tree was visualized with FigTree.

2.6. Locus-based phylogenetic analyses

To complement our SNP-based phylogenetic analyses, we also inferred phylogenetic trees based on the full sequence of the RAD loci, including both SNPs and invariant positions. For these datasets, we inferred phylogenetic relationships using both a maximum-likelihood and Bayesian approach. As the datasets containing the full-length sequences were substantially larger than the two SNP datasets, these analyses were considerably more computationally demanding. Consequently, all analyses described below were executed on the Argon high performance computing cluster at the University of Iowa.

To facilitate comparison of the SNP vs. RAD locus datasets, we first inferred trees using the maximum-likelihood approach in RAxML. To maintain consistency between these two analyses, we did not apply any additional partitions to our concatenated RAD locus dataset. Our analyses was performed as described above, using the GTRGAMMA evolutionary model, complete random starting trees for the 30% dataset, RAP trees for our 50% dataset, and the “-# autoMRE” data filtering option to determine the appropriate number of bootstrap replicates to use. Trees were visualized in FigTree.

For our Bayesian analysis, we selected the best evolutionary model for all loci using MrModelTest2 (v2.3; Nylander, 2004). We then implemented the Bayesian phylogenetic approach with MrBayes (v3.2.6; Huelsenbeck & Ronquist, 2001) on our 50% and 70% RAD loci datasets. For each dataset we simultaneously ran four Metropolis-coupled MCMC chains for one million generations under the best-fit evolutionary model, HLRT, determined by MrModelTest2. We used default parameters except a temperature of 0.06, which promoted the best chain mixing.
Of the four MCMC chains, we sampled one cold chain every 1,000 generations, and the other three were hot to allow exploration of parameter space. For each run, we discarded the first 25% of sampled trees as burn-in. We implemented two independent runs under these conditions and compared the output of the two runs when they had completed to ensure proper chain convergence. Trees were then visualized in FigTree.
3. Results

3.1. Illumina sequencing data

After demultiplexing and quality filtering our raw sequence reads, we recovered an average of 9.4±7.6 million reads per individual, which formed an average of 32,773±17,184 RAD loci in STACKS. This yielded an average sequencing depth of 154±84X. The number of RAD loci and SNPs recovered for each of the four datasets described in the methods are given in Table 2.

3.2. Inference of population structure

For both the 50% (A) and 30% (B) SNP datasets, the lowest C.V. error score was achieved by $K = 1$ with scores of 0.7 and 0.59, respectively, and thus our data best fit a model of $K = 1$, suggesting our data best fit a model of one population (Figure 2). However, since $K = 2$ performed only marginally worse in both datasets, we also investigated the patterns produced under a model with two populations for evidence of biologically meaningful population structure.

When looking at structure inferred under $K = 2$, both datasets recover similar results (Figure 3). In both cases, we find evidence of two distinct genetic clusters, corresponding to the Guiana Shield and Amazon Basin, respectively. In addition to this geographic split, three individuals (two from French Guiana and one from western Ecuador) have a distinct pattern of admixture, possibly indicative of the presence of an additional population, at both levels of missing data. Finally, two individuals from French Guiana collected from an atypical host (male flowers of *G. spinulosa*; indicated by squares in Figure 3), display evidence of mixed ancestry, and differ in their inferred ancestry compared to the other individuals collected in French Guiana.
Table 2. Number of RAD loci and SNPs missing from 30% or 50% of all 33 *Blepharoeura* sp. 10 individuals

<table>
<thead>
<tr>
<th></th>
<th># RAD loci</th>
<th># SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% missing data</td>
<td>5,905</td>
<td>4,674</td>
</tr>
<tr>
<td>50% missing data</td>
<td>12,961</td>
<td>12,085</td>
</tr>
</tbody>
</table>
Figure 2. C.V. error plots. Average 5-fold cross-validation error for the 50% (A) and 30% (B) datasets across 100 ADMIXTURE runs for each value of $K$. The error bars represent standard error above and below the average.
Figure 3. Population structure under a model of $K = 2$. To facilitate comparisons between the 50% (A) and 30% (B) datasets, both plots arrange the 33 Blepharoneura individuals by latitude/longitude within countries. Individuals are colored by their inferred ancestry proportions, with yellow corresponding to the Guiana Shield cluster, and grey to the Amazon Basin cluster. Symbols indicating host associations are consistent with those in Figure 1.
3.3. Phylogenetic analyses

Encouragingly, the topologies of all networks are largely concordant regardless of level of missing data, data type, and phylogenetic method utilized. Similar to our population structuring results, individuals collected from the Amazon Basin [Peru (P), Bolivia (B), Ecuador (E)] consistently form a distinct clade, as do individuals collected from the Guiana Shield [French Guiana (F), Suriname (S)]. In addition, three individuals from Trinidad, Venezuela, and western Ecuador (T1, V1, E1) consistently form a third clade, suggesting the existence of a third potential population within sp. 10. Although their exact placement varies, the two individuals from the French Guiana collected on male *G. spinulosa* (F4 and F5) generally do not form a clade with other individuals collected in the Guiana Shield. In most cases, they form their own distinct clade (Figures 4-6 and 8-11).
Figure 4. Phylogenetic network inferred with RAxML, based on concatenated RAD SNPs for the 50% dataset. Individuals are colored by country, and represented with symbols associated with their collection host.
Figure 5. Phylogenetic network inferred with RAxML, based on concatenated RAD SNPs for the 30% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 6. Phylogenetic network inferred with quartet-based analysis, as implemented in PAUP*, based on concatenated RAD SNPs for the 50% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 7. Phylogenetic network inferred with quartet-based analysis, as implemented in PAUP*, based on concatenated RAD SNPs for the 30% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 8. Phylogenetic network inferred with RAxML, based on concatenated RAD loci for the 50% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 9. Phylogenetic network inferred by RAxML, based on concatenated RAD loci for the 30% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 10. Phylogenetic network inferred using Bayesian inference, based on concatenated RAD loci for the 50% dataset. Symbols and colors are consistent with the keys found in Figure 4.
Figure 11. Phylogenetic network inferred using Bayesian inference, based on concatenated RAD loci for the 30% dataset. Symbols and colors are consistent with the keys found in Figure 4.
4. Discussion

Although the number of RAD loci and SNPs contained in either the datasets containing 50% or 30% missing data vary considerably, the amount of missing data does not qualitatively impact the population structuring results, nor the network topologies. In particular, both our population structuring and phylogenetic analyses support the existence of two distinct genetic clusters: one in the Amazon Basin, and the other in the Guiana Shield. These results are also consistent with previous work in the system (Ottens et al., 2017), which found evidence of population structure between the eastern and western portions of South America across several Blepharoneura species, including sp. 10.

Recently, demographic modeling of Blepharoneura sp 10 dated the divergence of the Amazon Basin and Guiana Shield populations to the late Pleistocene (Widmayer, 2018). The late Pleistocene was a time of relative instability, with repetitive cycles of glaciation and deglaciation occurring nearly world-wide (van der Hammen, 1974). Although most of South America remained unglaciated during the late Pleistocene, changes in climate and sea level altered the boundaries of the landmass and the distribution of habitat types from that we currently observe in South America. In particular, two major tropical rainforest refugia (Blepharoneura’s preferred habitat) are inferred to have existed in locations corresponding to the two major divisions we observe within Blepharoneura sp. 10 during the last glacial maximum (Varela et al. 2017). It is possible then that the differentiation we observe between these clusters is a result of historical isolation in these refugia.

In addition to these two major genetic clusters, we also found some support for an additional population consisting of singletons from Trinidad and Venezuela, as well as one individual from western Ecuador. In our population structuring analyses, these individuals all
display some amount of admixture, and cannot be confidently assigned to a single genetic cluster. Similarly, in our phylogenetic analyses, these individuals consistently form a distinct clade. Interestingly, a third, smaller region of tropical rainforest is inferred to have existed near the connection of Central and South America, which could have harbored this third population. A similar phenomenon is seen in *Heliconius* aldehyde studies, in which individuals from Panama, Trinidad, and western Ecuador lack consistent geographic structure (Turner et al., 1979).

The underlying genetic structure and major splits of many other Neotropical species, such as the *Heliconius* butterfly, may be related to Pleistocene changes in rainforest refugia. Similar to what is found in *Blepharoneura*, the two main groups of eastern and western *Heliconius* races consistently clade out in molecular analyses, and are found to have also diverged during the early Pleistocene (1.5-2 million years ago; Brower, 1994). Considering we see this occurrence in numerous South American biological systems, this may be a general pattern found in Neotropical systems and can explain some of the differentiation patterns found in *Blepharoneura* sp. 10. Together, our results suggest geographic location is a strong predictor of genetic differentiation within *Blepharoneura* sp. 10; and likely across most widely distributed *Blepharoneura* species. This differentiation found in *Blepharoneura* seems to have stemmed from historical isolation in multiple rainforest refugia during the last glacial maximum, but more detailed demographic analyses are needed to confirm this.

Although this has long been thought to be a common pattern in Neotropical species, more recent and accurate molecular clock analyses suggest that diversification and radiation events actually occurred later than this period of glaciation and deglaciation cycles. For example, in contrast to this widely accepted theory, Peres et al. (2015) determined that the range occupied by
the spider *Araneus omnicolor* remained stable and population growth remained unaffected despite the climactic fluctuations during the Pleistocene. Ramirez et al. (2010) also discovered that the major diversification of the *Euglossa* genus of bees did not occur until the Pliocene. The effect of Pleistocene climactic fluctuations on Neotropic species still remains not well understood, and it may be possible that these fluctuations do not have as much of an effect on species diversification as previously thought, and that diversification in *Blepharoneura* sp. 10 may be caused by other factors.

In addition to the observed geographic pattern, there is also some evidence that host plant use may drive some differentiation within *Blepharoneura* sp 10. The two individuals collected from male *G. spinulosa* flowers in French Guiana are consistently admixed in our population structuring analyses and fail to cluster with other individuals from the Guiana Shield in our phylogenetic analyses. Instead, these individuals variably cluster with individuals from the Amazon Basin, or with the individuals from Venezuela, Trinidad, and western Ecuador. One possible explanation for their inconsistent placement and apparent admixture may be that they truly represent an admixed population. An origin of this admixture may be a long distance migration event from the Amazon Basin to the Guiana Shield followed by subsequent hybridization with local individuals. This mixture of genetic material from divergent populations may have created new combinations of alleles, which in turn could facilitate their shift from female flowers (*Blepharoneura* sp. 10’s primary hosts) to male flowers, as genetic variability is needed for adaptive selection to work. The topologies of our phylogenetic networks support this idea, as these two individuals are found on relatively long branches, suggesting that they have undergone rapid evolutionary change, or that they have unique combinations of alleles.
Hybridization has long been thought to facilitate diversity and is found in many plant and animal species (Arnold, 1997). This is due to the fact that hybridization events greatly increase the amount of genetic variation in an individual (Burke & Arnold, 2001), as offspring with completely new combinations of alleles across hundreds of thousands of base pairs are created (Nolte & Tautz, 2010). A thoroughly studied biological system in which hybridization events allow a shift in ecological interactions is the *Helianthus* genus of sunflowers. When the genetic material of two parental species comes together to create unique combinations in a hybrid line, the hybrids are able to inhabit extreme environments that fall far outside of the tolerable range for the parental species (Rieseberg et al., 2003; Barton, 2008). Because hybridization events that produce offspring more fit to a wide variety of environments (Arnold & Martin, 2010) occur so frequently in nature, it could partially explain why some individuals in *Blepharoneura* sp. 10 have been able to change and survive off resources of an atypical host. Some examples do exist, though, of hybrid speciation that results from a host shift, instead of resulting in a host shift (Schwarz et al., 2005). Further analyses are needed to determine the exact order of events in *Blepharoneura* sp. 10.

What remains unclear, however, is if populations formed by historical isolation or putative host shifts will progress towards speciation, or remain only partially-isolated populations. If these populations are able to break free of each other through the development of reproductive barriers and restriction to gene flow, a speciation event will likely follow, but because *Blepharoneura* sp. 10 has not yet undergone diversification subsequent to its proposed split in the late Pleistocene, it is likely that gene flow is still occurring between the two populations.
A geographic barrier would indeed hinder the ability of these two populations to interbreed, but geographic isolation is not always strong enough to cause speciation, as shown by Thorpe et al (2010) with Anolis lizards separated throughout islands of an archipelago. An increase in evolutionary selective pressures between the different environments could also warrant a slow decrease in the amount of gene flow between the two populations, but the probability of when and if this occurs is hard to predict. Without the emergence of a barrier or mechanism that causes a significant decrease in gene flow between these two populations, it is unlikely that they will progress much further towards different species. If these populations fail to undergo ecological speciation, it is likely they will remain as they are, as partially isolated populations. We see this occurrence in a specific locality of Threespine Stickleback fish, in which separate populations develop divergent adaptations to different habitats but remain capable of interbreeding with each other via the absence of assortative mating (Räsänen et al, 2012). Gene flow among individuals of Blepharoneura sp. 10 across South America has had a very homogenizing effect, which makes it difficult to definitely say whether or not this species will diversify in the future.

Because we used a small pilot dataset of organisms that have no reference genome, there are some definite limitations in terms of what we could do and what we could interpret from our results. Similarly, our analyses are restricted to unrooted phylogenetic networks, as our dataset did not include an outgroup from another Blepharoneura species. Future work in this system will include more individuals of this and other Blepharoneura species, assembly of a robust genome, and inclusion of outgroups to facilitate more detailed population genomic and phylogenetic analyses. Rooting the trees may provide a more accurate representation of the number of genetic differences accumulated in each individual.
Additionally, because we used such a small sample size and limited host variability, there may be additional hierarchical population structure that we were not able to detect. In order to resolve finer scale patterns, in the future we may want to include a much larger cohort of samples spanning the entire width of their range in northern South America and from a larger amount of host variability. In addition, since our dataset has a significant sampling gap between the Guiana Shield and Amazon Basin, it is difficult to distinguish between true population structure and false structure inferred where there is actually continuous variation. However, the deep dating of divergence in demographic models and lack of isolation-by-distance within clusters suggests this is unlikely (see Widmayer 2018).

Regardless of which mechanisms have been most influential in the diversification of *Blepharoneura* sp. 10, we are starting to get a better understanding of how diversification occurs as a whole. There is no doubt that geographic barriers affect current and historical diversification events in South American systems (Colinvaux, 1989), but we have learned that other factors also play a significant role. There is clearly genetic distinction between eastern and western populations of sp. 10, which was likely caused by climactic variation in the late Pleistocene. The extent of influence Pleistocene climatic changes has on diversification, though, is still not well understood. *Blepharoneura* sp. 10 diversification also is influenced by host use, as individuals collected from an atypical host have significantly more genetic variation. These patterns that we have found in *Blepharoneura* sp. 10 are also being found in many other South American biological systems, and thus the scientific community is starting to get a better understanding of the mechanisms behind diversification of Neotropical species. Determining and understanding the mechanisms behind diversification of all species becomes critical when the need arises to
explain worldwide biodiversity patterns and how they contribute to the number of species on Earth (May, 1990; Wilson, 1992).
5. Literature Cited


