A continental-scale survey of *Wolbachia* infections in blue butterflies reveals evidence of interspecific transfer and invasion dynamics

Vivaswat Shastry[1], Katherine L. Bell[2], C. Alex Buerkle[3], James A. Fordyce[4], Matthew L. Forister[2], Zachariah Gompert[5], Sarah L. Lebeis[6], Lauren K. Lucas[5], Zach H. Marion[7] and Chris C. Nice[8]

[1] Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA

[2] Department of Biology, University of Nevada, Reno, NV 89557, USA

[3] Department of Botany, University of Wyoming, Laramie, WY 82071, USA

[4] Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

[5] Department of Biology, Utah State University, Logan, UT 84322, USA

[6] Department of Microbiology & Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

[7] Bio-Protection Research Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

[8] Department of Biology, Population and Conservation Biology, Texas State University, San Marcos, TX 78666, USA

Corresponding author: Vivaswat Shastry
Committee on Genetics, Genomics and Systems Biology
University of Chicago
Chicago, IL 60637, USA
vivaswat@uchicago.edu

Keywords: Wolbachia, Lycaeides, *infection acquisition, geography of infection, GBS data, host parasite interactions*

Running title: *Wolbachia* in *Lycaeides*

Abstract

Infections by maternally inherited bacterial endosymbionts, especially *Wolbachia*, are common in insects and other invertebrates but infection dynamics across species ranges are largely under studied. Specifically, we lack a broad understanding of the origin of *Wolbachia* infections in novel hosts, and the historical and geographical dynamics of infections that are critical for identifying the factors governing their spread. We used Genotype-by-Sequencing (GBS) data from previous population genomics studies for range-wide surveys of *Wolbachia* presence and genetic diversity in North American butterflies of the genus *Lycaeides*. As few as one sequence read identified by assembly to a *Wolbachia* reference genome provided high accuracy in detecting infections in host butterflies as determined by confirmatory PCR tests, and maximum accuracy was achieved with a threshold of only five sequence reads per host individual. Using this threshold, we detected *Wolbachia* in all but two of the 107 sampling localities spanning the continent, with infection frequencies within populations ranging from 0–100% of individuals, but with most localities having high infection frequencies (mean = 91% infection rate). Three major lineages of *Wolbachia* were identified as separate strains that appear to represent three separate invasions of *Lycaeides* butterflies by *Wolbachia*. Overall, we found extensive evidence for acquisition of *Wolbachia* through interspecific transfer between host lineages. Strain *wLyc*C was confined to a single butterfly taxon, hybrid lineages derived from it, and closely adjacent populations in other taxa. While the other two strains were detected throughout the rest of the continent, strain *wLyc*B almost always co-occurred with *wLyc*A. Our demographic modeling suggests *wLyc*B is a recent invasion. Within strain *wLyc*A, the two most frequent haplotypes are confined almost exclusively to separate butterfly taxa with haplotype A1 observed largely in *L. melissa* and haplotype A2 observed most often in *L. idas* localities, consistent with either cladogenic mode of infection acquisition from a common ancestor or by hybridization and accompanying mutation. More than one major *Wolbachia* strain was observed in 15 localities. These results demonstrate the utility of using resequencing data from hosts to quantify *Wolbachia* genetic variation and infection

frequency and provide evidence of multiple colonizations of novel hosts through hybridization between butterfly lineages and complex dynamics between *Wolbachia* strains.

Introduction

The endosymbiotic bacteria in the genus *Wolbachia* (Hertig & Wolbach, 1924; Hertig, 1936) have been studied for their phenotypic effects on their invertebrate hosts, mostly arthropods and nematodes (Yen & Barr, 1971; Charlat *et al.*, 2003; Moran *et al.*, 2008; Werren *et al.*, 2008; Kriesner *et al.*, 2013). The impacts on hosts include extraordinary reproductive manipulation as well as mutualistic interactions (Werren *et al.*, 2008). Reproductive manipulations include cytoplasmic incompatibility (mortality of host embryos when infected males mate with uninfected females), feminization, sex ratio distortion, and male killing. Mutualistic interactions are observed when *Wolbachia* infections protect hosts from viral attack (Teixeira *et al.*, 2008; Hedges *et al.*, 2008) or facilitate sequestration of vital nutrients (Brownlie *et al.*, 2009; Hosokawa *et al.*, 2010). These interactions have spurred development of evolutionary models to explain the persistence of infection within populations and the spread of infections across populations, lineages, and taxa (Turelli & Hoffmann, 1991; Turelli, 1994; Kriesner *et al.*, 2013, 2016). The manipulation of host biology has even been harnessed to control pest insect populations, including insects that vector human diseases (e.g., Kambris *et al.*, 2009; Hoffmann *et al.*, 2011; Iturbe-Ormaetxe *et al.*, 2011; Walker & Moreira, 2011; Ross *et al.*, 2019). However, despite nearly a half a century of research on the phenotypic effects of *Wolbachia* on their hosts, we have a relatively poor understanding of 1) how *Wolbachia* infects novel hosts, and 2) the historical biogeography of infection dynamics within and among host lineages.

Wolbachia infections occur in more than half of insect species (Werren *et al.*, 1995; Hilgenboecker *et al.*, 2008; Zug & Hammerstein, 2015; Weinert *et al.*, 2015; Bailly-Bechet *et al.*, 2017) and acquisition by novel hosts can occur in multiple ways. Cladogenic acquisition, also known as co-speciation or co-divergence, occurs when infections are acquired from an ancestral lineage, resulting in sister taxa sharing common *Wolbachia* strains or genotypes (Cooper *et al.*, 2019; Sanaei *et al.*, 2021). Introgressive acquisition occurs through reproduc-

tive exchange between lineages (i.e., via hybrid formation) and constitute host shifts (Cooper *et al.*, 2019; Sanaei *et al.*, 2021). Alternatively, horizontal transfer might result from parasitoid (Stevens *et al.*, 2001; Heath *et al.*, 1999; Duron *et al.*, 2010; Gehrer & Vorburger, 2012; Gupta *et al.*, 2020) or ectoparasite attack (Hoy & Jeyaprakash, 2005; Le Clec'h *et al.*, 2013; Gupta *et al.*, 2020), or possibly through predation and other food sources (Huigens *et al.*, 2000, 2004; Gerth *et al.*, 2013). Evidence for cladogenic acquisition is sparse (Turelli *et al.*, 2018; Raychoudhury *et al.*, 2009; Gerth *et al.*, 2013) and requires phylogenetic information from *Wolbachia* and hosts. Distinguishing between introgressive acquisition and the various pathways of horizontal transfer not involving reproductive interaction can be accomplished by comparisons of divergence times of *Wolbachia* and mitochondrial DNA (Conner *et al.*, 2017; Turelli *et al.*, 2018).

Once acquired, the dynamics of *Wolbachia* prevalence in a population are presumably governed by the phenotypic effects on hosts, host immune responses to infections, coevolution, and the fidelity of *Wolbachia* transmission from host mother to offspring (Turelli, 1994; Weeks *et al.*, 2007; Jaenike, 2009; Hoffmann, 2020; Sanaei *et al.*, 2021). For example, models of cytoplasmic incompatibility-inducing strains, where infected females experience fitness effects that are dependent on the frequency of infected males predict that prevalence within a host population can exhibit decreases in frequency when infections are rare, but rapidly increase when infection frequency is above an unstable threshold frequency. Such "bistable" infection dynamics could produce variation in prevalence among host populations that depend on initial infection frequencies where some populations have low to non-existent infections, whereas others might be fixed, or nearly fixed, for the infection (Barton & Turelli, 2011; Turelli & Hoffmann, 1991; Kriesner *et al.*, 2013). Alternatively, strains with positive, frequency-independent fitness effects on their hosts are predicted to increase in frequency within populations and spread spatially (Barton & Turelli, 2011; Kriesner *et al.*, 2013) much like alleles under positive selection. Range-wide surveys of infection prevalence across species ranges provide critical information for understanding the spread and maintenance of infec-

tions (Hague *et al.*, 2021). Such data can also be used to estimate the patterns of strain distribution among host species to inform models of evolution and transmission of *Wolbachia* (Cooper *et al.*, 2019; Turelli *et al.*, 2018; Kriesner *et al.*, 2013). Further, temporal sampling could also be incorporated to investigate the factors that govern the spatial spread of infections (Riegler *et al.*, 2005; Kriesner *et al.*, 2013).

Despite a robust theoretical foundation for *Wolbachia* infection acquisition by novel hosts and evolutionary dynamics, detailed range-wide surveys of *Wolbachia* prevalence in natural populations and species have been conducted in a limited number of hosts (e.g., Shoemaker *et al.*, 2000, 2003a,b; Narita *et al.*, 2006; Nunes *et al.*, 2008; Baldo *et al.*, 2008; Turelli *et al.*, 2018; Schuler *et al.*, 2016, 2018; Walker *et al.*, 2021; Hague *et al.*, 2021) or ecological communities (Gupta *et al.*, 2020). Historically, *Wolbachia* infections have been assayed using PCR-based methods targeting *Wolbachia* 16S rDNA genes or other *Wolbachia*-specific markers such as the Multilocus Strain Typing (MLST) loci (Turelli & Hoffmann, 1995; Werren & Windsor, 2000; Baldo *et al.*, 2006, 2008). Presence of *Wolbachia* in a host individual is indicated by an amplified band on agarose gels and strain identification can be performed by sequencing the MLST loci (Baldo *et al.*, 2006, 2008). However, PCR-based assays can be time-consuming, especially for surveys of large numbers of individual hosts, and are subject to false-positive errors from contamination of samples and false-negative errors from failed PCR reactions, among other problems. While various methods have been developed to minimize them (e.g. Turelli & Hoffmann, 1995; Nice *et al.*, 2009), these errors can present challenges for range-wide surveys of infections in hosts.

A promising and inexpensive approach for such surveys uses *Wolbachia* resequencing data from phylogeographic and population genomics studies of host species (Richardson *et al.*, 2012; Signor, 2017; Pascar & Chandler, 2018; Hinojosa *et al.*, 2020; Arif *et al.*, 2021). Here, we pursue a bioinformatics approach similar to these previous studies and use a large population genomics data set from multiple species of *Lycaeides* butterflies to highlight the potential of this approach for estimating infection prevalence and to support inferences of

modes of acquisition and histories of infections.

Lycaeides butterflies colonized North America through Beringia approximately 2–4 million years ago (Gompert *et al.*, 2008a; Vila *et al.*, 2011). This colonization was followed by a period of diversification that included extensive admixture among lineages (Nice *et al.*, 2013; Gompert *et al.*, 2014). We recognize five lineages that correspond to nominal species (Fig. 1 and Table 1; Gompert *et al.*, 2014), including: *L. idas*, which also occurs in the Palearctic, and four North American endemics: *L. melissa*, *L. anna*, *L. ricei* and *L. samuelis*. *Lycaeides samuelis* (formerly *L. melissa samuelis*) is known as the Karner blue butterfly and is a federally listed endangered species (Black & Vaughan, 2005; Forister *et al.*, 2011). We also recognize three distinct lineages within *L. melissa*: *L. melissa*-East, *L. melissa*-Rockies and *L. melissa*-West (Chaturvedi *et al.*, 2018). In addition to the nominal species, there are several admixed lineages that we refer to as hybrid lineages. These occur in several mountain ranges of the western United States. Putative ancient hybrid lineages formed from admixture between *L. melissa* and *L. anna* occur in the Sierra Nevada and in the White Mountains of California and Nevada (Gompert *et al.*, 2006a; Nice *et al.*, 2013; Gompert *et al.*, 2014). In the vicinity of Jackson, Wyoming in the Grand Tetons and Yellowstone area of the Rocky Mountains, we find populations that exhibit admixture between *L. melissa* and *L. idas* that we refer to as the Jackson hybrid lineage (Gompert *et al.*, 2012, 2013, 2014; Chaturvedi *et al.*, 2020). Hybrid lineages in the Warner Mountains of northeastern California, the Jarbidge Range in northern Nevada, and Steens Mountain in southeastern Oregon, have complex ancestry potentially from *L. melissa*, *L. anna* and *L. idas* (Gompert *et al.* 2014).

PCR-based surveys have demonstrated that North American *Lycaeides* harbor *Wolbachia* infections (Gompert *et al.*, 2008b; Nice *et al.*, 2009). For example, populations of *L. samuelis*, Karner blue butterflies, in the western portion of their range in Wisconsin were found to be nearly entirely infected (near 100% prevalence). These populations also possessed a mitochondrial haplotype identical to a haplotype found in *L. melissa* (the proposed source of the infection), but distinct from haplotypes found in the eastern portion of their range

(east of Lake Michigan) (Nice *et al.*, 2009). However, surveys of *Wolbachia* in *Lycaeides* have been limited in terms of geography and butterfly taxonomy. Here, we expand our survey to provide a continent-wide view of *Wolbachia* diversity using sequence reads from population genomics studies of *Lycaeides*.

The data considered here are GBS sequence reads from 2,377 butterflies of the genus *Lycaeides* from 107 localities in North America sampled from 1996 to 2018 (Table 1, Supplementary Table 1). These data were generated for several projects investigating patterns of differentiation and admixture across North America (Gompert *et al.*, 2014), genomic changes during shifts to novel host plants (Gompert *et al.*, 2015; Chaturvedi *et al.*, 2018), and comparisons of genomic architecture between ancient hybrid lineages and a contemporary hybrid zone (Chaturvedi *et al.*, 2020). A chromosome-level reference genome for *L. melissa* has been assembled to facilitate comparative genomic studies (Chaturvedi *et al.*, 2018, 2020).

We used *Wolbachia* sequence reads found among Genotyping-by-Sequencing (GBS) reads from North American *Lycaeides* butterflies to address the following questions: 1) how does variation in detection thresholds of sequence depth and sequence length influence *Wolbachia* infection frequency estimation? 2) how does the frequency of infection (prevalence) vary across host populations, lineages, and geography? and 3) how are *Wolbachia* genotypes and groups of genotypes (which we equate to strain types) distributed across geography and host taxonomy? We use the answers to these questions to construct hypotheses about the history and biogeography of infections in *Lycaeides*. We also discuss the opportunities and limitations to using *Wolbachia* reads present in resequencing data as an inexpensive tool for understanding *Wolbachia* dynamics in natural populations. We also argue that similar data from other host taxa might contribute to the growing understanding of the evolution and history of *Wolbachia*-host interactions.

Materials and Methods

Sequencing of *Lycaeides* individuals

We extracted genomic DNA, generated GBS libraries, and sequenced these libraries following the methods described in Parchman *et al.* (2012), Gompert *et al.* (2014), and Chaturvedi *et al.* (2018). Briefly, genomic DNA was extracted from thoracic tissue for all specimens and purified using Qiagen's DNeasy Blood and Tissue kit (Qiagen Inc). Genomic DNA was digested with restriction enzymes EcoRI and MseI. Adapters, including a unique 8–10bp sequence barcode and the Illumina primer sequences, were ligated to the fragmented DNA with T4 DNA ligase. Adaptor and primer sequences are provided in Supplementary Table 2. We then PCR-amplified the fragment libraries with standard Illumina PCR primers. Amplified libraries were then pooled and size-selected (300–450bp) with a BluePippin. The GBS libraries were sequenced across several lanes of Illumina HiSeq 2500 or HiSeq 4000 (100bp, single-end reads) by the Genome Sequencing and Analysis Facility at the University of Texas (Austin, TX).

Obtaining *Wolbachia* sequence reads from host GBS reads

Though we knew from PCR-based surveys that *Lycaeides* butterflies harbor *Wolbachia* (Gompert *et al.*, 2008a; Nice *et al.*, 2009), we had very little information about strain types or even the diversity of strains that might be encountered in a broader survey. Preliminary assemblies of sequence reads from *Lycaeides* to different publicly available *Wolbachia* genomes revealed variation in number of assembled reads across localities and reference genomes (data not presented). We interpreted this as a possible indication that butterflies from different localities, or taxa, harbored a diversity of *Wolbachia* strains and that different reference genomes might yield better assemblies for some strains and therefore also some host localities or taxa.

Given this possibility, we explored several assembly strategies for creating reference genomes *in-silico*. Our first assembly was performed by concatenating three *Wolbachia* reference genomes representing each of supergroups A, B and F (Ramírez-Puebla *et al.*, 2015). Reference genomes are available for supergroups A–F. *Wolbachia* supergroups A and B are commonly found in insects. Supergroup F is found in both insects and nematodes, while, among supergroups with representative *Wolbachia* reference genomes, supergroups C, D and E are found exclusively in nematodes (Ramírez-Puebla *et al.*, 2015). Thus, this concatenated genome we constructed represented the most likely supergroups that might be observed in *Lycaeides* butterflies. The three representative reference genomes came from *Wolbachia* in *Drosophila melanogaster* (*wMel*, supergroup A; Wu *et al.*, 2004), *Aedes albopictus* (*wAlbB*, supergroup B; Mavingui *et al.*, 2012), and *Cimex lectularius* (*wCle*, supergroup F; Nikoh *et al.*, 2014).

Because nuclear integration of *Wolbachia* genes into the host genome (e.g. Nikoh *et al.*, 2008; Choi *et al.*, 2015) is possible, we mapped putative *Wolbachia* reads identified by assembly to the concatenated reference genomes described above to the *L. melissa* reference genome (Chaturvedi *et al.*, 2018) (details in Supplementary Table 3). Upon querying the location of the mapped *Wolbachia* reads in the host genome, we found that *all* the reads mapped to one of the smaller scaffolds (Scaffold 1260, 1.62 Mbp) out of the 1,651 scaffolds in the *L. melissa* genome, and not from any of the larger chromosomal level (23 autosomes and Z sex chromosome) scaffolds. Based on the length (similar to size of other *Wolbachia* genomes) and mapping metrics of this region, we believe this scaffold to be the genome of *Wolbachia* infecting the host butterfly individual used in the genome assembly. We then pursued a second assembly in which we used this Scaffold 1260 as a species-specific reference *Wolbachia* genome for further analysis.

Lastly, we used a pan-genome approach (Tettelin *et al.*, 2005; Vernikos *et al.*, 2015) to build a reference representing the super set of genes from the above *Wolbachia* reference genomes. Here we used the supergroup A, B and F genomes described above plus the

Scaffold 1260 from the *Lycaeides* reference genome. The pan-genome was constructed by first annotating the representative *Wolbachia* genomes using `prokka` (version 1.14.6, Seemann, 2014) to convert the `fasta` files to `gff` format (using the 'Moderate' parameters in `https://github.com/tseemann/prokka`, last accessed Jul 13, 2022), and then combining the files to produce a reference pan-genome using the `Roary` (version 3.13.0, Page *et al.*, 2015) and `GNU Parallel` softwares (Tange, 2011). Genes with paralogs and genes with less than 98% BLASTp percentage identity with each other were removed from the pan-genome. Finally, the pan-genome contained 11, 114, and 4,547 genes (total: 4,672 genes) present in three, two and one of the four constituent genomes respectively. The pan-genome was also considerably larger at approximately 3.25 Mbp (as expected), compared to Scaffold 1260 (1.62 Mbp) and the concatenated reference (2.8 Mbp).

In all three iterations (i.e., using the concatenated reference, Scaffold 1260 from the *Lycaeides* reference genome, or the pan-genome), reads were aligned using `bowtie2` software (all aligned reads reported using `-a --al --no-unal`, version 2.3.4.2, Langmead & Salzberg, 2012). The mapped reads from each of the 2,377 individuals were output as `sam` files to allow for easy parsing and analysis downstream. Multiple previous studies (e.g., Richardson *et al.*, 2012; Signor, 2017; Scholz *et al.*, 2020) show that an approach similar to the above is effective in not only retrieving large amounts of endosymbiont genomic data from host reads but also conducting population-level analyses on the extracted endosymbiont lineages. There were minor differences in the metrics of the intermediate bioinformatics analyses (e.g., number of reads, etc., listed in Supplementary Table 3) depending on the reference genome used, but we found very similar results in detecting infected individuals (see below) and in the final construction of the haplotype networks, identification of strains and in the geographic patterns of genetic variation (see below) across these assemblies. As a result, our subsequent analyses were based on the pan-genome reference assembly as this contains a super set of our genes from all assemblies. Details about the analyses with the other reference genomes are presented in the Supplementary Material (for instance, gene annotations for the pseudo-

haplotype in the Scaffold 1260 reference presented in Supplementary Table 4).

Detecting infection from the mapped *Wolbachia* reads

We first quantified the number of mapped reads and the length of mapped reads in an individual's `sam` file from the pan-genome assembly as two metrics for detecting infected individuals. We then examined how various minimum thresholds of these two metrics affected the classification of host individuals as infected and compared the results to a previous PCR-based study of a subset of the current individuals (128 out of 2,377) from Nice *et al.* (2009). To collect these metrics, we used `samtools view -F 2432` (skipping secondary alignments, version 1.12, Li *et al.*, 2009) to determine the length (in base pairs, bp) and number of reads of each unique individual alignment for reads filtered to have a mapping quality of greater than 20 (less than a 1% chance of error, as is standard in typical pipelines). A similar type of bioinformatics approach has been previously used successfully by Pascar & Chandler (2018) to detect *Wolbachia* infection in various arthropod species. We deviate from previous purely bioinformatics studies by choosing a more appropriate threshold (for *Lycaeides*) for infection that maximizes concordance of infection status with results from the previous PCR-based amplification study in these same butterfly species.

Quantifying genetic diversity in *Wolbachia* strains

The individual `sam` files were each compressed into `bam` files (using `samtools`) for more efficient downstream analyses. We then performed variant calling and genotyping on the sorted and indexed `bam` files from the previous step using the `bcftools mpileup` command (skipping indels) followed by the `bcftools call` and `view` commands (version 1.9, Li, 2011) to produce a raw `vcf` file across all 2,377 individuals using the pangenome reference genome described above. We ignored indels, assumed a ploidal level of one (haploid) and retained only bi-alleleic sites (`--ploidy 1 --variants-only -m2 -M2 -v snps`). The choice to employ a

haploid model was based on preliminary analyses from a diploid model. Given the existence of more than one major *Wolbachia* strain and sympatry among strains in some instances (see Results), it is possible that individual host butterflies could contain multiple infections (i.e., a single individual hosting two or more *Wolbachia* strains). However, models for variant calling with higher ploidy (for instance, a diploid model that might be more appropriate for multiple infections) compromised our ability to call variants as haplotypes because phasing of alleles at multiple sites was not possible. Therefore, we employed the haploid model to produce useful haplotypic data. This undoubtedly prevented discovery of additional haplotypes in individuals with multiple infections, but did produce population genetic data for those individuals with single infections. The raw `vcf` file was then filtered to only keep sites with a maximum missingness of 25% using `vcftools` (version 0.1.14, Danecek *et al.*, 2011). The final `vcf` file contained 115 SNPs and 2,377 individuals in total, as a result of our conservative filtering.

The alleles in each individual from this `vcf` file could now be regarded as representing *Wolbachia* haplotypes. However, to minimize uncertainty in the haplotypic data, we again filtered the data by retaining individuals with **no** missing data across variant sites (i.e., individuals with no missing data had at least 1 read of mapping quality greater than 20 of either the reference or the alternative allele at every site). We retained 1,277 individuals (out of 2,113 infected individuals, see Results) with haplotypes of length 115bp.

We clustered the individual haplotypes using a statistical parsimony network approach (Templeton *et al.*, 1992; Crandall *et al.*, 2000) using the `haplotypes` (version 1.1.2, Aktas, 2020) package in R with a parsimony threshold of 95%. All analyses in R was performed on version 4.0.3 (R Core Team, 2021). As a complimentary approach, we performed a Principal Coordinates Analysis (PCoA) on the matrix of pairwise sequence distances calculated with the `haplotypes` package and using the `prcomp` function in R. Based on these analyses, we identified three major groups of haplotypes that we consider as distinct *Wolbachia* strains (see Results). Strain types, and haplotypes within strains, were then mapped onto the

geographical and taxonomic distributions of the host butterflies. Thus, the distribution of *Wolbachia* strains and haplotypes (chloropleth maps produced using `tmap v3.3-1`, Tennekes, 2018) were examined in the context of the biogeography of their hosts and used to construct hypotheses about the origin and dynamics of infection within *Lycaeides* butterflies.

Reconstructing demographic history of *Wolbachia* strains

Lastly, we investigated the demographic history, specifically, changes in effective population size through time, for each of the three major strains (see Results) to understand *Wolbachia* population dynamics. We created a `NEXUS` file of all haplotypes from each of the three major strains (see Results) and used `BEAST v2.6.3` to estimate Bayesian Skyline Plots (BSP, Drummond *et al.*, 2005). This method fits a piece-wise linear function to the estimated population size as calculated from coalescent rates across the sequence. A single long chain, total of 75 million steps with a burn-in of 50 million steps, thinned every 50,000 steps for *wLyc*A and a total of 50 million steps with a burn-in of 10 million steps, thinned every 50,000 steps for *wLyc*B and *wLyc*C, due to the large number of individuals and parameters in *wLyc*A, was run. We ran a coalescent Bayesian skyline analysis with a HKY site model (Hasegawa *et al.*, 1985) with a strict clock and a uniform prior on the clock rate. The full settings in the `BEAUti` files are presented in the Supplementary Material. Convergence to a posterior distribution was assessed based on visualizations of the trace plots and calculation of effective sample sizes (ESS) of the posterior distribution for each network using `Tracer v1.7.1` (Rambaut *et al.*, 2018), which was also used to obtain uncorrelated parameter estimates from the sampling distribution.

Results

Genotype-by-Sequencing data for *Lycaeides* individuals

For the 2,377 *Lycaeides* individuals sequenced, a total of 3,727,714,988 sequence reads were generated (mean = 1,568,244 per individual, median = 1,363,955 per individual). From the *Wolbachia* mapping protocol described in the Methods section and on filtering for reads with mapping quality (MAPQ) greater than 20, we obtained approximately 8.75 million reads spread across all individuals, with a median of approximately 3,500 mapped reads per individual and more than 90% of the reads having lengths greater than 80 bp. The total *Wolbachia* reads comprise approximately 0.2% of all sequence reads. The distribution of mapped read lengths is shown in Supplementary Fig. 1.

Detecting infection from the mapped *Wolbachia* reads

We set our detection threshold for infection in individual butterflies at a minimum of 5 reads of $>$ 80bp (with the maximal length being 87bp). We found that results from this threshold matched very well with results from a previous PCR amplification study (Nice *et al.*, 2009), with a 96.9% accuracy rate (i.e., concordance with PCR-based results, Fig. 2). A threshold read length of 80bp was chosen since this was very close to the largest possible read from an individual, and would act as a stringent threshold for infection detection. We also found that $>$ 90% of mapped reads had lengths greater than 80bp (see Supplementary Fig. 1). Similarly, we chose 5 reads as our threshold sequence depth because this threshold minimized error with comparison to PCR tests. We found that by increasing the threshold number of mapped reads we were increasing our false negative rate for classification by declaring putatively infected individuals (based on PCR tests) as being uninfected. This type of approach results in a sharp drop in the accuracy as we increase the threshold beyond a read depth of 600 since fewer individuals are classified as being infected (due to the stringent threshold) and

therefore, increase the false negative error in our comparison. The 5 reads threshold provided a good balance between the false positive and false negative error rates (Fig. 2). However, we note that the PCR-based amplification studies are also prone to inaccuracies that could affect our accuracy estimates.

The numbers of infected individuals were not substantially changed by varying the minimum number of reads required to diagnose infected individuals for most localities. The exceptions where prevalence did vary with different thresholds were localities for Karner Blue butterflies in the eastern portion of their range (Indiana Dunes (5), Allegan (6), Saratoga (7)) and several of the Sierra Nevada hybrid lineage localities (98-99, 101-105). (Note: when referring to specific localities, we include the site number(s) from Table 1 and Fig. 1 in parentheses following the locality names.) In these localities, raising the minimum number of reads substantially reduced the number of infected individuals detected. Supplementary Table 5 presents the numbers of infected individuals using thresholds of a minimum of 1×, 5× and 20× reads, and Supplementary Fig. 2 provides a detailed examination of the relationship between minimum number of reads and read lengths on the percentage of infected individuals detected across all individuals.

Based on the threshold of a minimum of five reads of at least 80bp (Supplementary Fig. 2), we found that a majority of *Lycaeides* localities had infection frequencies that exceeded 90% of individuals, with 85 of the 107 sampled localities showing greater than 90% (with 64 localities having infection frequencies of 100%) (Fig. 3, Table 1). In populations where we observed variation for infection (i.e. infection frequencies not zero or one), 90.6% of females and 86.4% of males were infected (population treated as a random effect, $\chi^2 = 4.62$, df=1, p-value = 0.032). At the species or lineage level, most infections rates are greater than 94% (Table 2). The exceptions included *L. samuelis* localities in the eastern portion of their range (5-7) (infection rates: 0-0.5%), one population of *L. ricei* from the Marble Mts. in California (38) (infection rate: 58%), a small number of *L. melissa* populations, mostly in the western Great Basin (43, 73, 75, 76) (infection rates: 80-88%), one population of the

hybrid lineages in the Jackson area (at Swift Creek (92) (infection rate: 75%)), in the Sierra Nevada (98, 99, 101-105) (infection rates: 15-75%) and in the White Mountains (106, 107) (infection rates: 87-89%) (Table 1).

Quantifying genetic diversity in *Wolbachia* strains

The filtered `vcf` file with 115 variable sites and 1,277 individuals was used for population genetic analyses. Based on a haplotype network analysis with 95% statistical parsimony and PCoA of pairwise distances among haplotypes, we found that 1,267 out of 1,277 genotyped individuals carried *Wolbachia* haplotypes from one of three major haplotype networks (Fig. 4, Supplementary Fig. 3) that correspond to three clusters of haplotypes in our ordination of haplotypes (Fig. 5). We consider these networks as distinct *Wolbachia* strain types with individual haplotypes within networks representing mutational variation within strains (referred to hereafter as *wLyc*A, *wLyc*B and *wLyc*C) (Table 1, Supplementary Table 1). Each of these strains included between 3 and 44 distinct but closely related haplotypes (Fig. 4). The strains were substantially divergent from one another with mean pairwise divergences between strains ranging from 11.4% to 37.4% (Table 3). In addition, the diversity of *Wolbachia* strains and haplotypes within butterfly populations varied widely (Supplementary Table 1). Butterfly sampling localities ranged from localities that contained a single *Wolbachia* haplotype to localities with a maximum of 15 haplotypes (at Girl Farm (70)). The highest strain diversities were observed in the western Great Basin *L. melissa* populations and in some of the localities of hybrid lineages of *Lycaeides* (Table 1 , Supplementary Table 1 and Supplementary Fig. 4).

Wolbachia strain *wLyc*A was observed in 992 individuals and was the most frequent strain. Among the 19 haplotypes within strain *wLyc*A, haplotypes A1 and A2 were observed in 936 individuals (94% of individuals with *wLyc*A haplotypes). Though these two haplotypes were differentiated by a single mutational step (Fig. 4), they were mostly observed in different butterfly taxa. The A1 haplotype was found almost exclusively in *L. melissa*, while A2 was

limited to *L. idas* (Tables 1, 2). The exceptions include all three disjunct *L. idas* localities sampled in Colorado (26-28) where A1 was observed; A1 was also observed in the four *L. samuelis* localities sampled in Wisconsin (1-4) (results that match earlier PCR-based surveys (Nice *et al.*, 2009)); A2 was observed in the *L. melissa* population at Albion Meadows (65), in the Jackson hybrids (82-92) and hybrid lineages in the Warner Mountains (93-94), Jarbidge Mountains (97) and at Steens Mountain (95) (notably not in the hybrid lineages in the Sierra Nevada and White Mountains in California and Nevada (98-107) for which *L. anna* is the maternal parent) (Table 2). Both A1 and A2 were also observed in the contemporary hybrid zone between *L. melissa* and *L. idas* at Dubois (85) (Chaturvedi *et al.*, 2020), in the *L. ricei* population at Cave Lake (37) and in two *L. melissa* localities in the Rockies (61, 64) (Supplementary Fig. 5).

Strain *wLyc*B was observed in 103 individuals and included 44 haplotypes. Strain *wLyc*B haplotypes occurred most frequently in the populations of *L. melissa* in the western Great Basin (68-78) ("L. melissa - West" in Table 1 and Supplementary Table 1). Haplotypes B1 and B10 were the most common *wLyc*B haplotypes in these western populations. The other haplotypes occur in low frequency in these *L. melissa* West populations and at County Line (106), part of the hybrid lineage in the White Mountains, and in two *L. idas* populations (21, 25), one population of *L. melissa* East (56), the Big Lake (33) population of *L. ricei*, and three populations of *L. idas* (Spruce Barley (30), Garnet Peak (21) and Hayden Valley (25); Supplementary Fig. 6).

Strain *wLyc*C was observed in 172 individuals and included three haplotypes. Strain *wLyc*C haplotypes were confined to *L. anna* populations (8-13) and hybrids in the Sierra Nevada and White Mountains (99, 101-107). These hybrids have mixed ancestry from both *L. melissa* and *L. anna* and the latter is presumed to be the maternal lineage based on patterns of mtDNA variation (Gompert *et al.*, 2006a). The two exceptions for the distribution of *wLyc*C haplotypes was their presence in the Shovel Creek, CA (39) and Marble Mt.s (38) *L. ricei* populations, which are the southern-most sampled *L. ricei* localities and adjacent

to the range of *L. anna* (Supplementary Fig. 7).

For all three major strains, we found distinct right-skewed frequency distributions with one to six haplotypes observed in the majority of individuals and the remaining haplotypes were found in relatively few individuals, often spread over extensive areas (Fig. 4, Table 1). The remaining ten *Lycaeides* individuals that did not possess *Wolbachia* haplotypes from strains *wLyc*A, *wLyc*B or *wLyc*C contained very rare haplotypes assigned to seven rare strains (*wLyc*D-*wLyc*J) that were observed as singletons, five (D2, E1, F1, H1 and H2) in four localities in the western Great Basin (70,71,73,78) (*L. melissa* West), one (D1) in the County Line (106) hybrid population, two (G1 and G2) in the Marble Mountains (38), one (I1) at Wheatland (*L. melissa* Rockies, 63), and one (J1) at Yuba Gap (*L. anna*, 9) (Table 1, Supplementary Fig. 2).

Reconstructing demographic history of *Wolbachia* strains

Based on our analysis of demographic history across the haplotypes within each of the three major strains, we find different patterns for each strain in the past (Fig. 4). We found well-mixed trace plots for all three strains and ESS values of about 200 for strain *wLyc*A and *wLyc*C, and about 400 for strain *wLyc*B (all three above the recommended threshold for independent samples from the `BEAST2` manual). Strain *wLyc*A (which contains mostly the *L. melissa* and *L. idas* individuals) shows a constant scaled population size of 0.02 stretching into the very distant past. Strain *wLyc*B (which includes individuals from the western Great Basin (68-78), the hybrid lineages in the White Mountains in California (106) and Jackson, Wyoming area (83-84)) seems to have existed at much higher population sizes ($\sim 2.5\times$ population size of strain *wLyc*A) in the distant past, but has experienced a growth phase starting 0.003 time units in the past and has grown up to $\sim 4\times$ its previous size since then. Strain *wLyc*C (which is observed in the *L. anna* individuals and adjacent localities) has a very small and constant population size (roughly $0.1\times$ of the the other two strains) stretching into the distant past. Based on the tree event times presented in

Supplementary Fig. 8, we observe that both *wLyc*A and *wLyc*B strains have undergone population size changes in the recent past whereas strain *wLyc*C shows the highest spike at time 0, indicating that population size has been relatively constant over previous time periods. The time units are measured in substitutions and we assume equal rates across the strains to aid in interpretation.

Discussion

We used a bioinformatics approach for detecting *Wolbachia* infection from GBS reads of 2,377 *Lycaeides* butterflies and validated the results from this approach by comparison with PCR-based analyses of a small subset of the host individuals (Nice *et al.*, 2009). Using a threshold of a minimum of five reads of at least 80bp, we found that most individuals were infected (2,117 out of 2,377 surveyed) and 105 out of 107 localities contained infected individuals. Infection prevalences within locality samples ranged from 0–100% of individuals infected with a mean infection prevalence per locality of 91% infected individuals. Population genetic analyses of *Wolbachia* haplotype data provided relatively detailed phylogeographic information on three major *Wolbachia* strains that infect *Lycaeides* butterflies in North America. Examination of the geographic and host-taxonomic distributions of *Wolbachia* strains revealed extensive sharing of strains between populations and lineages of *Lycaeides* which represents evidence for introgressive acquisition (Tables 1, 2). Coalescent-based demographic inferences also provided evidence that one of the major strains has had a recent and dramatic increase in effective population size and might currently be invading and possibly displacing another strain.

Varying the threshold minimum sequence length had little effect on detecting infected individuals because the vast majority of sequence reads were greater than 80bp in length (Supplementary Fig. 1). While the threshold of a minimum of 5 (five) reads provided the greatest accuracy (based on comparisons to PCR surveys), varying the minimum number

of reads threshold had a limited impact on estimated infection frequencies except in 10 localities (5-7, 98-99, 101-105, see Supplementary Table 2 and Supplemetary Table 5). In these localities, increasing the minimum reads threshold substantially reduced our estimate of prevalence of infected individuals. Three of these localities occur in the eastern portion of the range of the Karner Blue butterfly (*L. samuelis*) (5-7), but Karner blue populations in the western portion of the range (1-4) do not exhibit the same reduction in estimated prevalence with increasing minimum reads threshold. Similarly, the other localities that show the decline in numbers of infected individuals with increasing minimum reads threshold occur in the hybrid lineage of *Lycaeides* in the Sierra Nevada (98-99, 101-105), yet other hybrid lineages do not show a similar pattern. It is not immediately obvious why these localities differed in their apparent sensitivity to the minimum reads threshold. The overall number of sequence reads per individual could affect the probability of detection, but while the eastern Karner localities have lower median number of reads compared to the total set of 2,377 individuals, the Sierran hybrid populations have more reads per individual than the overall median (median number of sequence reads: eastern Karners: 1,078,622, Sierran hybrids: 1,810,680, overall: 1,359,589). Alternatively, it is possible that there is variation in *Wolbachia* densities within individuals among localities that influences detection probability (Unckless *et al.*, 2009; Hague *et al.*, 2021; Shropshire *et al.*, 2021). While we cannot explain this observation at present, it suggests that variation in *Wolbachia* infection densities in host tissues might be an important consideration when mining resequencing data for evidence of endosymbiont infection. Variation among host taxa might require careful inspection of these thresholds. In the absence of corroborating PCR-based data on infection status, we recommend examining a range of thresholds to understand how these affect the probability of detection. It is also possible that more sophisticated statistical modeling that accounts for uncertainty created by variation in numbers of sequence reads, and possibly variation in *Wolbachia* densities, could improve the probability of detecting infections.

Population genetic analyses of *Wolbachia* infections in the *Lycaeides* system facilitated

inference of infection history. We do not know where or how the three major *Wolbachia* strains (*wLyc*A, *wLyc*B and *wLyc*C) were ultimately acquired by *Lycaeides* in North America. Analysis of *Wolbachia* infections from *Lycaeides* from Europe and Asia, or from associated parasites or parasitoids, might shed light on the origins of North American infections. However, our survey of geographic patterns of population genetic variation in combination with inference of demographic histories of the three major strains suggest that transmission of infection within North American *Lycaeides* butterflies occurred primarily through introgressive acquisition. We provide an overview of these patterns.

The comparison of demographic histories of each strain, as coalescent effective population sizes ($N_e\mu$), is facilitated by previous evidence for constant *Wolbachia* substitution rates over long timescales (Cooper *et al.*, 2019). The demographic history of strain *wLyc*A reveals a relatively constant population size over time, and the geographic and taxonomic distribution of strain *wLyc*A haplotypes is possibly consistent with either a cladogenic mode or an introgressive mode of acquisition. The two most frequent haplotypes in *wLyc*A (A1 and A2) exhibit just one mutational difference (Fig. 4), yet A1 is largely confined to *L. melissa* individuals and A2 is found almost exclusively in *L. idas* individuals (Table 1, Fig. 6). Exceptions to this pattern include hybrid lineages with either *L. melissa* or *L. idas* ancestry, or ancestry from both species (i.e., in the Jackson, Wyoming area (82-92), the contemporary hybrid zone between *L. melissa* and *L. idas* at Dubois (85), or localities at or near range boundaries, such as Cave Lake (37)). The confinement of these haplotypes largely within two *Lycaeides* species seems compatible with the hypothesis of cladogenic acquisition in the ancient past through a common ancestral lineage of *L. idas* and *L. melissa*, followed by independent divergence of the two lineages. Alternatively, the distribution of haplotypes A1 and A2 might be consistent with introgression from one of the species into the other accompanied by mutation. Further, the exceptions to the distributional pattern (e.g., hybrid lineages and a hybrid zone) appear to be examples of introgressive acquisition of strain *wLyc*A haplotypes outside of *L. melissa* and *L. idas*. Thus, there is perhaps more support

for introgressive acquisition of *wLyc*A haplotypes, though cladogenetic aquisition cannot be ruled out. Evidence for multiple modes of *Wolbachia* transmission in natural populations is also found in the *Drosophila* (Cooper *et al.*, 2019) and *Nasonia* (Raychoudhury *et al.*, 2010) species complexes.

A similar demographic history of constant population size over time is seen in strain *wLyc*C, though the estimated population size of *wLyc*C is very much smaller than the other strains (Fig. 4). Strain *wLyc*C haplotypes are confined to *L. anna* populations and the hybrid lineages in the Sierra Nevada and White Mountains (98-107) for which *L. anna* is the presumed maternal lineage (Table 1, Fig. 6) (Gompert *et al.*, 2006a). The exceptions include two localities where strain *wLyc*C haplotypes were observed, both of which lie on the boundary between the ranges of *L. anna* and *L. ricei* at the Marble Mountains (38) and Shovel Creek (39). Thus, as with strain *wLyc*A haplotypes, *wLyc*C haplotypes appear to have spread to a limited extent outside of a *Lycaeides* species range via introgression among lineages, specifically in this case from *L. anna* to nearby populations of *L. ricei*.

The phylogeography of strain *wLyc*B is different compared to the other two strains. This is the least frequently observed strain over all and the majority of *wLyc*B haplotypes were observed in the western Great Basin in populations of *L. melissa* (68-78) (Table 1, Fig. 6). In these locations, multiple *wLyc*B haplotypes are commonly observed along with A haplotypes. In fact, *wLyc*B haplotypes were observed without accompanying A haplotypes in only three locations (Verdi Tracks (73), Deer Mt. Road (76) and Gardnerville (78)). However, *wLyc*B haplotypes were observed in other widely distributed places and other *Lycaeides* taxa including: *L. idas* in Alaska (30) and Montana (21), *L. ricei* at Cave Lake (37), *L. melissa* in central Nevada (56), in the Jackson hybrid lineage (84), the hybrid lineage in the White Mountains of California (106), and the putative hybrid lineage at Hinkley in northern Nevada (96). The relative rarity of this strain, its recent population expansion (Fig. 4), coupled with its presence almost exclusively with *wLyc*A across different host species points to an introgressive mode of acquisition. Strain *wLyc*B haplotypes appear to be

invading localities that already contain infections of strain *wLyc*A. Such a mode of acquisition will lead to the presence of multiple *Wolbachia* infections or haplotypes from different strains segregating in the same population and hence, an enriched genetic diversity of *Wolbachia* in these populations (Supplementary Fig. 4, Supplementary Table 1). The concentration of strain *wLyc*B haplotypes in *Lycaeides* localities in the western Great Basin, and the resulting high haplotype diversity there, suggests that this area is where the invasion of strain *wLyc*B began. There is weak evidence for the hypothesis that strain *wLyc*B is invading from one locality, Verdi Crystal (71), that was sampled over multiple years and strain *wLyc*B appears to have increased in frequency from 2011 - 2018 (Supplementary Table 6). The studies from which these GBS data were obtained were not designed to assay *Wolbachia* or for temporal comparisons, and we lack statistical power to fully test this hypothesis without further sampling. The host butterflies at these localities have colonized alfalfa (*Medicago sativa*) relatively recently (Forister *et al.*, 2020a,b), probably as one of three or more independent colonizations of alfalfa (Chaturvedi *et al.*, 2018), and probably within the last 200 years (400–600 butterfly generations Chaturvedi *et al.*, 2018; Forister *et al.*, 2020a,b). So, it is possible that we are tracking the effect of host population expansion in the demographic history of strain *wLyc*B as it is impossible to disentangle the two histories without more information on host demography and quantification of *Wolbachia* titer levels. Thus, it seems that the invasion of a novel *Wolbachia* strain is occurring while the butterfly host is switching to a novel host plant. Whether there is any connection between these parallel host switches is an open question.

At a continental scale, the nominal species or lineages of *Lycaeides* butterflies each contained a dominant (most frequently observed) major strain (Table 2). Some lineages shared major strains. For example, *L. melissa* and *L. samuelis* shared strain *wLyc*A (specifically haplotype A1). This pattern is consistent with interspecific transfer from *L. melissa* to *L. samuelis* (Gompert *et al.*, 2006b; Nice *et al.*, 2009). Beyond their specific dominant strains, most lineages also contained other "minor" strains that were dominant in other lineages

but observed at lower frequency in the focal lineage (Table 2). These minor strains were commonly observed at range margins and are consistent with limited interspecific transfer. Hybrid lineages were observed to be infected by major strains associated with their putative maternal parental lineage. The Sierra/Whites hybrid lineage is infected with *wLyc*C as is the inferred maternal parent *L. anna*. Similarly, the Jackson hybrid lineage is infected with *wLyc*A, specifically haplotype A2, as is its maternal parental lineage *L. idas* (Table 2). Taken together, these observations illustrate considerable interspecific transfer of *Wolbachia* strains among host lineages.

The distribution of *Wolbachia* strains in *Lycaeides* butterflies is paralleled by geographical patterns of mitochondrial DNA variation observed in previous studies of these butterflies (Nice *et al.*, 2005; Gompert *et al.*, 2008a,b). Because *Wolbachia* infections and mitochondrial DNA (mtDNA) are maternally inherited, they are commonly observed to be in linkage disequilibrium (Hurst & Jiggins, 2005; Jiggins, 2003; Turelli & Hoffmann, 1991; Turelli *et al.*, 1992). Direct comparisons are not possible because even where sampling localities overlap with the current study, those older studies of mtDNA variation used different individuals that were not sequenced for this study. Nevertheless, the presence of three major *Wolbachia* strains discovered here parallels the three major mtDNA lineages discovered in *Lycaeides*. For example, using mitochondrial sequences of the cytochrome oxidase I (COI) and cytochrome oxidase II (COII) genes, Gompert *et al.* (2008b) found three mitochondrial lineages. One lineage (lineage III from Gompert *et al.* (2008b)) was widely distributed across space and butterfly taxonomy that corresponds to the distribution of *wLyc*A here. Another mtDNA lineage (lineage II) co-occurred with the first lineage and was detected in populations of *L. melissa* from the western Great Basin and from the hybrid population in the White Mt.s (County Line, 106), corresponding to the distribution of *wLyc*B. Lastly, the third mtDNA lineage (lineage I) was observed in *L. anna*, hybrid lineages derived from *L. anna* in the Sierra Nevada and adjacent *L. ricei* localities, corresponding to *wLyc*C. The close geographical correspondence of major *Wolbachia* lineages observed here and previous mtDNA

haplotype distributions suggest that the expected disequilibrium between *Wolbachia* strains and mtDNA can be detected using GBS data.

Our survey of *Wolbachia* infection frequencies and genetic variation using GBS data from host *Lycaeides* butterflies suggests that this approach could be applied in other systems. Given the quantity of resequencing data generated recently, it might be possible to rapidly survey *Wolbachia* and other endosymbiont infections in a wide variety of host organisms and answer broad questions about the history, geography and mode of acquisition of infections. However, resequencing methods do not specifically target *Wolbachia* genomes and there exist several limitations. The sequence reads from *Lycaeides* GBS data did not map to any of the multi-locus sequence typing (MLST) loci (Baldo *et al.*, 2006, 2008) and it seems unlikely that GBS data in general will overlap the MLST loci. Thus, it will be impossible to identify conventionally-designated strains (or possibly even *Wolbachia* supergroups) and connect studies phylogenetically from surveys of GBS data without further sequencing. Additionally, the stochasticity inherent in the methods for resequencing data, combined with the sparseness of endosymbiont sequence reads from host organisms, presents some challenges. Stochasticity arising from library preparation and from the sequencing of these multiplexed genomic libraries, among other possible sources of stochasticity, creates variation in sequence depth across fragments and individuals. This variation can contribute to false negatives for infection detection. Given variation in sequencing effort across studies, we note that the threshold for infection detection (here we used a minimum of 5 sequence reads) will need to be carefully examined for each investigation. False positives from GBS data seem less likely than false negatives compared to PCR-based methods for infection detection, though contamination of samples is an important consideration for both PCR-based and GBS survey methods. The usefulness of resequencing data for population genetics investigations of endosymbionts will be facilitated by the development of new methods for detecting infection and for genotyping that can, for example, more fully account for uncertainty and accommodate the possibility of multiple infections within individuals. Despite these limitations, the

use of resequencing data can cheaply and relatively easily facilitate surveys of endosymbiont infection and population genetics.

Data Accessibility

DNA-sequence data have been deposited in the NCBI SRA with accession codes PRJNA246037, PRJNA577236, PRJNA432816 and PRJNA862870. Scripts for analysis are uploaded to `https://github.com/VivaswatS/wolbachia_lycaeides.git`.

Author Contributions

VS and CCN conceived and designed the study. All authors collected data. VS analysed the data. VS and CCN drafted the initial version of the manuscript. All authors contributed to later versions of the manuscript.

Acknowledgments

We thank J. Ott and M. Turelli for critical comments and discussion. We thank N. Anthony, G. Gelembiuk, D. Raterman, C. O'Brien, and M. Amaral for collections of *Lycaeides samuelis* under USFWS permit PRT842392. J. Oliver, K. Prudic, J. Jahner, and A. Epeset helped with collections. We also thank Dr. Hidenori Tachida and two anonymous reviewers for help in improving the manuscript. We also thank members of the University of Chicago population genetics community for providing recommendations on appropriate bioinformatics tools to use. Computing was performed in the Teton Computing Environment at the Advanced Research Computing Center (University of Wyoming, `https://doi.org/10.15786/M2FY47`) and at the LEAP High Performance Computing Cluster at Texas State University. This work was supported by National Science Foundation grant DEB-1638793 to MLF, DEB-1638768

to ZG, DEB-1638773 to CCN, DEB-1638922 to JAF, and DEB-1638602 to CAB; MLF was additionally supported by a Trevor James McMinn professorship.

References

Aktas C (2020) *haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony*. R package version 1.1.2.

Arif S, Gerth M, Hone-Millard WG, Nunes MD, Dapporto L, Shreeve TG (2021) Evidence for multiple colonisations and *Wolbachia* infections shaping the genetic structure of the widespread butterfly *Polyommatus icarus* in the British Isles. *Molecular Ecology*, **30**, 5196–5213.

Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot MF, Charlat S (2017) How long does *Wolbachia* remain on board? *Molecular Biology and Evolutiontion*, **34**, 1183–1193.

Baldo L, Ayoub NA, Hayashi CY, Russell JA, Stahlhut JK, Werren JH (2008) Insight into the routes of *Wolbachia* invasion: high levels of horizontal transfer in the spider genus *Agelenopsis* revealed by *Wolbachia* strain and mitochondrial DNA diversity. *Molecular Ecology*, **17**, 557–569.

Baldo L, Dunning Hotopp JC, Jolley KA, *et al.* (2006) Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Applied and environmental microbiology*, **72**, 7098–7110.

Barton NH, Turelli M (2011) Spatial waves of advance with bistable dynamics: cytoplasmic and genetic analogues of allee effects. *The American Naturalist*, **178**, E48–E75.

Black S, Vaughan D (2005) Species profile: *Lycaeides melissa samuelis*. In: *Red List of*

671 *Pollinator Insects of North America*, The Xerces Society for Invertebrate Conservation
672 Portland, OR.

673 Bouckaert R, Heled J, Kühnert D, *et al.* (2014) BEAST 2: a software platform for Bayesian
674 evolutionary analysis. *PLoS Comput Biol*, **10**, e1003537.

675 Brownlie JC, Cass BN, Riegler M, *et al.* (2009) Evidence for metabolic provisioning by a
676 common invertebrate endosymbiont, *Wolbachia pipientis*, during periods of nutritional
677 stress. *PLoS Pathog*, **5**, e1000368.

678 Charlat S, Hurst GD, Merçot H (2003) Evolutionary consequences of *Wolbachia* infections.
679 *TRENDS in Genetics*, **19**, 217–223.

680 Chaturvedi S, Lucas LK, Buerkle CA, *et al.* (2020) Recent hybrids recapitulate ancient
681 hybrid outcomes. *Nature Communications*, **11**, 1–15.

682 Chaturvedi S, Lucas LK, Nice CC, Fordyce JA, Forister ML, Gompert Z (2018) The pre-
683 dictability of genomic changes underlying a recent host shift in Melissa blue butterflies.
684 *Molecular Ecology*, **27**, 2651–2666.

685 Choi JY, Bubnell JE, Aquadro CF (2015) Population genomics of infectious and integrated
686 *Wolbachia pipientis* genomes in *Drosophila ananassae*. *Genome Biology and Evolution*, **7**,
687 2362–2382.

688 Conner WR, Blaxter ML, Anfora G, Ometto L, Rota-Stabelli O, Turelli M (2017)
689 Genome comparisons indicate recent transfer of wRi-like *Wolbachia* between sister species
690 *Drosophila suzukii* and *D. subpulchrella*. *Ecology and Evolutiontion*, **7**, 9391–9404.

691 Cooper BS, Vanderpool D, Conner WR, Matute DR, Turelli M (2019) *Wolbachia* acquisition
692 by *Drosophila yakuba*-clade hosts and transfer of incompatibility loci between distantly
693 related *Wolbachia*. *Genetics*, **212**, 1399–1419.

694 Crandall MCDPK, Clement M, Posada D (2000) TCS: a computer program to estimate gene
695 genealogies. *Molecular Ecology*, **9**, 1657–1660.

696 Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and vcftools. *Bioin-*
697 *formatics*, **27**, 2156–2158.

698 Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of
699 past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**,
700 1185–1192.

701 Duron O, Wilkes TE, Hurst GD (2010) Interspecific transmission of a male-killing bacterium
702 on an ecological timescale. *Ecology Letters*, **13**, 1139–1148.

703 Forister M, Philbin C, Marion Z, *et al.* (2020a) Predicting patch occupancy reveals the
704 complexity of host range expansion. *Science Advances*, **6**, eabc6852.

705 Forister ML, Gompert Z, Fordyce JA, Nice CC (2011) After 60 years, an answer to the
706 question: what is the karner blue butterfly? *Biology Letters*, **7**, 399–402.

707 Forister ML, Yoon SA, Philbin CS, *et al.* (2020b) Caterpillars on a phytochemical landscape:
708 The case of alfalfa and the Melissa blue butterfly. *Ecology and Evolution*, **10**, 4362–4374.

709 Gehrer L, Vorburger C (2012) Parasitoids as vectors of facultative bacterial endosymbionts
710 in aphids. *Biology Letters*, **8**, 613–615.

711 Gerth M, Röthe J, Bleidorn C (2013) Tracing horizontal *Wolbachia* movements among bees
712 (*Anthophila*): a combined approach using multilocus sequence typing data and host phy-
713 logeny. *Molecular Ecology*, **22**, 6149–6162.

714 Gompert Z, Fordyce JA, Forister ML, Nice CC (2008a) Recent colonization and radiation
715 of North American *Lycaeides* (*Plebejus*) inferred from mtDNA. *Molecular Phylogenetics*
716 *and Evolution*, **48**, 481–490.

717 Gompert Z, Fordyce JA, Forister ML, Shapiro AM, Nice CC (2006a) Homoploid hybrid
718 speciation in an extreme habitat. *Science*, **314**, 1923–1925.

719 Gompert Z, Forister ML, Fordyce JA, Nice CC (2008b) Widespread mito-nuclear discordance
720 with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Molecular*
721 *Ecology*, **17**, 5231–5244.

722 Gompert Z, Jahner JP, Scholl CF, *et al.* (2015) The evolution of novel host use is unlikely to
723 be constrained by trade-offs or a lack of genetic variation. *Molecular Ecology*, **24**, 2777–
724 2793.

725 Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (2014) Admixture
726 and the organization of genetic diversity in a butterfly species complex revealed through
727 common and rare genetic variants. *Molecular Ecology*, **23**, 4555–4573.

728 Gompert Z, Lucas LK, Nice CC, Fordyce JA, Alex Buerkle C, Forister ML (2013) Geograph-
729 ically multifarious phenotypic divergence during speciation. *Ecology and Evolutiontion*, **3**,
730 595–613.

731 Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic
732 regions with a history of divergent selection affect fitness of hybrids between two butterfly
733 species. *Evolution: International Journal of Organic Evolution*, **66**, 2167–2181.

734 Gompert Z, Nice CC, Fordyce JA, Forister ML, Shapiro AM (2006b) Identifying units for
735 conservation using molecular systematics: the cautionary tale of the karner blue butterfly.
736 *Molecular Ecology*, **15**, 1759–1768.

737 Gupta M, Kaur R, Gupta A, Raychoudhury R (2020) Are ecological communities the seat
738 of endosymbiont horizontal transfer and diversification? A case study with soil arthropod
739 community. *Authorea Preprints*.

740 Hague MT, Shropshire J, Caldwell C, *et al.* (2021) Temperature effects on cellular host-
741 microbe interactions explain continent-wide endosymbiont prevalences. *Current Biology*,
742 **in press**.

743 Hasegawa M, Kishino H, Yano Ta (1985) Dating of the human-ape splitting by a molecular
744 clock of mitochondrial dna. *Journal of molecular evolution*, **22**, 160–174.

745 Heath BD, Butcher RD, Whitfield WG, Hubbard SF (1999) Horizontal transfer of *Wolbachia*
746 between phylogenetically distant insect species by a naturally occurring mechanism. *Cur-*
747 *rent Biology*, **9**, 313–316.

748 Hedges LM, Brownlie JC, O'Neill SL, Johnson KN (2008) *Wolbachia* and virus protection
749 in insects. *Science*, **322**, 702–702.

750 Hertig M (1936) The *rickettsia*, *Wolbachia pipientis* (gen. et sp. n.) and associated inclusions
751 of the mosquito, *Culex pipiens*. *Parasitology*, **28**, 453–486.

752 Hertig M, Wolbach SB (1924) Studies on rickettsia-like micro-organisms in insects. *The*
753 *Journal of medical research*, **44**, 329.

754 Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH (2008) How many
755 species are infected with *Wolbachia*?–a statistical analysis of current data. *FEMS micro-*
756 *biology letters*, **281**, 215–220.

757 Hinojosa JC, Koubínová D, Dincă V, *et al.* (2020) Rapid colour shift by reproductive char-
758 acter displacement in cupido butterflies. *Molecular Ecology*, **29**, 4942–4955.

759 Hoffmann A (2020) *Wolbachia*. *Current Biology*, **30**, R1113–R1114.

760 Hoffmann AA, Montgomery B, Popovici J, *et al.* (2011) Successful establishment of *Wol-*
761 *bachia* in *Aedes* populations to suppress dengue transmission. *Nature*, **476**, 454–457.

762 Hosokawa T, Koga R, Kikuchi Y, Meng XY, Fukatsu T (2010) *Wolbachia* as a bacteriocyte-
763 associated nutritional mutualist. *Proceedings of the National Academy of Sciences*, **107**,
764 769–774.

765 Hoy MA, Jeyaprakash A (2005) Microbial diversity in the predatory mite *Metaseiulus oc-*
766 *cidentalis* (acari: *Phytoseiidae*) and its prey, *Tetranychus urticae* (acari: *Tetranychidae*).
767 *Biological Control*, **32**, 427–441.

768 Huigens M, De Almeida R, Boons P, Luck R, Stouthamer R (2004) Natural interspecific and
769 intraspecific horizontal transfer of parthenogenesis–inducing *Wolbachia* in *Trichogramma*
770 wasps. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **271**,
771 509–515.

772 Huigens M, Luck R, Klaassen R, Maas M, Timmermans M, Stouthamer R (2000) Infectious
773 parthenogenesis. *Nature*, **405**, 178–179.

774 Hurst GD, Jiggins FM (2005) Problems with mitochondrial DNA as a marker in population,
775 phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings*
776 *of the Royal Society B: Biological Sciences*, **272**, 1525–1534.

777 Iturbe-Ormaetxe I, Walker T, O'Neill SL (2011) *Wolbachia* and the biological control of
778 mosquito-borne disease. *EMBO reports*, **12**, 508–518.

779 Jaenike J (2009) Coupled population dynamics of endosymbionts within and between hosts.
780 *Oikos*, **118**, 353–362.

781 Jiggins FM (2003) Male-killing *Wolbachia* and mitochondrial DNA: selective sweeps, hybrid
782 introgression and parasite population dynamics. *Genetics*, **164**, 5–12.

783 Kambris Z, Cook PE, Phuc HK, Sinkins SP (2009) Immune activation by life-shortening
784 *Wolbachia* and reduced filarial competence in mosquitoes. *Science*, **326**, 134–136.

785 Kriesner P, Conner WR, Weeks AR, Turelli M, Hoffmann AA (2016) Persistence of a *Wol-*
786 *bachia* infection frequency cline in *Drosophila melanogaster* and the possible role of repro-
787 ductive dormancy. *Evolution*, **70**, 979–997.

788 Kriesner P, Hoffmann AA, Lee SF, Turelli M, Weeks AR (2013) Rapid sequential spread of
789 two *Wolbachia* variants in *Drosophila simulans*. *PLoS pathogens*, **9**, e1003607.

790 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nature Methods*,
791 **9**, 357.

792 Le Clec'h W, Chevalier FD, Genty L, Bertaux J, Bouchon D, Sicard M (2013) Cannibalism
793 and predation as paths for horizontal passage of *Wolbachia* between terrestrial isopods.
794 *PloS ONE*, **8**, e60232.

795 Li H (2011) A statistical framework for snp calling, mutation discovery, association mapping
796 and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**,
797 2987–2993.

798 Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and
799 samtools. *Bioinformatics*, **25**, 2078–2079.

800 Mavingui P, Moro CV, Tran-Van V, *et al.* (2012) Whole-genome sequence of *Wolbachia*
801 strain wAlbB, an endosymbiont of tiger mosquito vector *Aedes albopictus*.

802 Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bac-
803 terial symbionts. *Annual review of genetics*, **42**, 165–190.

804 Narita S, Nomura M, Kato Y, Fukatsu T (2006) Genetic structure of sibling butterfly species
805 affected by *Wolbachia* infection sweep: evolutionary and biogeographical implications.
806 *Molecular Ecology*, **15**, 1095–1108.

807 Nice CC, Anthony N, Gelembiuk G, Raterman D, ffrench constant R (2005) The history

808 and geography of diversification within the butterfly genus *Lycaeides* in North America.
809 *Molecular Ecology*, **14**, 1741–1754.

810 Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA (2013) Hybrid speci-
811 ation and independent evolution in lineages of alpine butterflies. *Evolution: International*
812 *Journal of Organic Evolution*, **67**, 1055–1068.

813 Nice CC, Gompert Z, Forister ML, Fordyce JA (2009) An unseen foe in arthropod conser-
814 vation efforts: the case of *Wolbachia* infections in the Karner blue butterfly. *Biological*
815 *Conservation*, **142**, 3137–3146.

816 Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T (2014) Evolutionary
817 origin of insect–*Wolbachia* nutritional mutualism. *Proceedings of the National Academy of*
818 *Sciences*, **111**, 10257–10262.

819 Nikoh N, Tanaka K, Shibata F, *et al.* (2008) *Wolbachia* genome integrated in an insect
820 chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome*
821 *research*, **18**, 272–280.

822 Nunes MD, Nolte V, Schlötterer C (2008) Nonrandom *Wolbachia* infection status of
823 *Drosophila melanogaster* strains with different mtDNA haplotypes. *Molecular Biology and*
824 *Evolutiontion*, **25**, 2493–2498.

825 Page AJ, Cummins CA, Hunt M, *et al.* (2015) Roary: rapid large-scale prokaryote pan
826 genome analysis. *Bioinformatics*, **31**, 3691–3693.

827 Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012)
828 Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecol-*
829 *ogy*, **21**, 2991–3005.

830 Pascar J, Chandler CH (2018) A bioinformatics approach to identifying *Wolbachia* infections
831 in arthropods. *PeerJ*, **6**, e5486.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, **67**, 901.

Ramírez-Puebla ST, Servín-Garcidueñas LE, Ormeño-Orrillo E, *et al.* (2015) Species in *Wolbachia*? Proposal for the designation of 'candidatus *Wolbachia bourtzisii*','candidatus *Wolbachia onchocercicola*','candidatus *Wolbachia blaxteri*','candidatus *Wolbachia brugii*','candidatus *Wolbachia taylori*','candidatus *Wolbachia collembolicola*'and 'candidatus *Wolbachia multihospitum*'for the different species within *Wolbachia* supergroups. *Systematic and applied microbiology*, **38**, 390–399.

Raychoudhury R, Baldo L, Oliveira DC, Werren JH (2009) Modes of acquisition of *Wolbachia*: horizontal transfer, hybrid introgression, and codivergence in the *Nasonia* species complex. *Evolution: International Journal of Organic Evolution*, **63**, 165–183.

Raychoudhury R, Grillenberger BK, Gadau J, *et al.* (2010) Phylogeography of *Nasonia vitripennis* (*Hymenoptera*) indicates a mitochondrial–*Wolbachia* sweep in north america. *Heredity*, **104**, 318–326.

Richardson MF, Weinert LA, Welch JJ, *et al.* (2012) Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003129.

Riegler M, Sidhu M, Miller WJ, O'Neill SL (2005) Evidence for a global *Wolbachia* replacement in *Drosophila melanogaster*. *Current Biology*, **15**, 1428–1433.

Ross PA, Turelli M, Hoffmann AA (2019) Evolutionary ecology of *Wolbachia* releases for disease control. *Annual Review of Genetics*, **53**, 93–116.

Sanaei E, Charlat S, Engelstädter J (2021) *Wolbachia* host shifts: routes, mechanisms, constraints and evolutionary consequences. *Biological Reviews*, **96**, 433–453.

Scholz M, Albanese D, Tuohy K, Donati C, Segata N, Rota-Stabelli O (2020) Large scale genome reconstructions illuminate *Wolbachia* evolution. *Nature Communications*, **11**, 1–11.

Schuler H, Egan SP, Hood GR, Busbee RW, Driscoe AL, Ott JR (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**, 1–15.

Schuler H, Köppler K, Daxböck-Horvath S, *et al.* (2016) The hitchhiker's guide to europe: the infection dynamics of an ongoing *Wolbachia* invasion and mitochondrial selective sweep in *Rhagoletis cerasi*. *Molecular Ecology*, **25**, 1595–1609.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Shoemaker D, Keller G, Ross KG (2003a) Effects of *Wolbachia* on mtDNA variation in two fire ant species. *Molecular Ecology*, **12**, 1757–1771.

Shoemaker DD, Ahrens M, Sheill L, Mescher M, Keller L, Ross KG (2003b) Distribution and prevalence of *Wolbachia* infections in native populations of the fire ant *Solenopsis invicta* (*Hymenoptera*: *Formicidae*). *Environmental Entomology*, **32**, 1329–1336.

Shoemaker DD, Ross KG, Keller L, Vargo E, Werren JH (2000) *Wolbachia* infections in native and introduced populations of fire ants (*Solenopsis* spp.). *Insect Molecular Biology*, **9**, 661–673.

Shropshire JD, Hamant E, Cooper BS (2021) Male age and *Wolbachia* dynamics: Investigating how fast and why bacterial densities and cytoplasmic incompatibility strengths vary. *mBio*, **12**, e02998–21.

Signor S (2017) Population genomics of *Wolbachia* and mtDNA in *Drosophila simulans* from California. *Scientific Reports*, **7**, 1–11.

880 Stevens L, Giordano R, Fialho RF (2001) Male-killing, nematode infections, bacteriophage
881 infection, and virulence of cytoplasmic bacteria in the genus *Wolbachia*. *Annual Review of*
882 *Ecology and Systematics*, **32**, 519–545.

883 Tange O (2011) GNU parallel - the command-line power tool. *The USENIX Magazine*, **36**,
884 42–47.

885 Teixeira L, Ferreira Á, Ashburner M (2008) The bacterial symbiont *Wolbachia* induces re-
886 sistance to rna viral infections in *Drosophila melanogaster*. *PLoS Biol*, **6**, e1000002.

887 Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations
888 with haplotypes inferred from restriction endonuclease mapping and dna sequence data.
889 iii. cladogram estimation. *Genetics*, **132**, 619–633.

890 Tennekes M (2018) tmap: Thematic maps in R. *Journal of Statistical Software*, **84**, 1–39.

891 Tettelin H, Masignani V, Cieslewicz MJ, *et al.* (2005) Genome analysis of multiple pathogenic
892 isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceed-*
893 *ings of the National Academy of Sciences*, **102**, 13950–13955.

894 Turelli M (1994) Evolution of incompatibility-inducing microbes and their hosts. *Evolution*,
895 **48**, 1500–1513.

896 Turelli M, Cooper BS, Richardson KM, *et al.* (2018) Rapid global spread of wRi-like *Wol-*
897 *bachia* across multiple *Drosophila*. *Current Biology*, **28**, 963–971.

898 Turelli M, Hoffmann A, McKechnie SW (1992) Dynamics of cytoplasmic incompatibility and
899 mtDNA variation in natural *Drosophila simulans* populations. *Genetics*, **132**, 713–723.

900 Turelli M, Hoffmann AA (1991) Rapid spread of an inherited incompatibility factor in Cal-
901 ifornia *Drosophila*. *Nature*, **353**, 440–442.

902 Turelli M, Hoffmann AA (1995) Cytoplasmic incompatibility in *Drosophila simulans*: dy-
903 namics and parameter estimates from natural populations. *Genetics*, **140**, 1319–1338.

Unckless RL, Boelio LM, Herren JK, Jaenike J (2009) *Wolbachia* as populations within individual insects: causes and consequences of density variation in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 2805–2811.

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. *Current opinion in microbiology*, **23**, 148–154.

Vila R, Bell CD, Macniven R, *et al.* (2011) Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 2737–2744.

Walker T, Moreira LA (2011) Can *Wolbachia* be used to control malaria? *Memórias do Instituto Oswaldo Cruz*, **106**, 212–217.

Walker T, Quek S, Jeffries CL, *et al.* (2021) Stable high-density and maternally inherited *Wolbachia* infections in *Anopheles moucheti* and *Anopheles demeilloni* mosquitoes. *Current Biology*, **31**, 2310–2320.

Weeks AR, Turelli M, Harcombe WR, Reynolds KT, Hoffmann AA (2007) From parasite to mutualist: rapid evolution of *Wolbachia* in natural populations of *Drosophila*. *PLoS Biol*, **5**, e114.

Weinert LA, Araujo-Jnr EV, Ahmed MZ, Welch JJ (2015) The incidence of bacterial endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20150249.

Werren JH, Baldo L, Clark ME (2008) *Wolbachia*: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, **6**, 741–751.

Werren JH, Windsor DM (2000) *Wolbachia* infection frequencies in insects: Evidence of a global equilibrium? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **267**, 1277–1285.

928 Werren JH, Zhang W, Guo LR (1995) Evolution and phylogeny of *Wolbachia*: reproductive
929 parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological*
930 *Sciences*, **261**, 55–63.

931 Wu M, Sun LV, Vamathevan J, *et al.* (2004) Phylogenomics of the reproductive parasite
932 *Wolbachia pipientis wMel*: a streamlined genome overrun by mobile genetic elements.
933 *PLoS Biol*, **2**, e69.

934 Yen JH, Barr AR (1971) New hypothesis of the cause of cytoplasmic incompatibility in culex
935 pipiens l. *Nature*, **232**, 657–658.

936 Zug R, Hammerstein P (2015) Bad guys turned nice? A critical assessment of *Wolbachia*
937 mutualisms in arthropod hosts. *Biological Reviews*, **90**, 89–111.

Figures and Tables

Table 1: Sample information for 107 *Lycaeides* butterfly collection localities. Locality numbers, locality names, nominal species designations (see text for details), number of individuals sampled, number of infected individuals detected, data source and *Wolbachia* haplotypes (Fig. 4) are provided. Infected individuals were identified using a threshold of a minimum of five sequence reads of at least 80bp in length. The data source column indicates previously published sequence data (G = Gompert *et al.* (2014), C = Chaturvedi *et al.* (2018) or sequence data presented here for the first time (present).

#	Locality	Nominal Species	n	# Infected	Data	Wolbchia haplotype
1	Fish Lake	*L. samuelis*	20	20	present	A1(14)
2	Eau Claire	*L. samuelis*	22	21	present	A1(2)A8(1)
3	Black River	*L. samuelis*	17	17	present	A1(14)
4	Fort McCoy	*L. samuelis*	23	23	present	A1(20)A10(1)
5	Indiana Dunes	*L. samuelis*	21	1	present	
6	Allegan	*L. samuelis*	30	0	present	
7	Saratoga Spr.s	*L. samuelis*	27	0	present	
8	Fall Cr	*L. anna*	20	20	G	C1(11)
9	Yuba Gap	*L. anna*	20	20	G	C1(14)J1(1)
10	Castle Pk	*L. anna*	18	16	G	C1(9)
11	Donner Pass	*L. anna*	18	17	G	C1(4)
12	Marlette Lk	*L. anna*	19	19	present	C1(9)
13	Leek Spr.s	*L. anna*	20	20	G	C1(18)
14	Cottonwood	*L. idas*	25	25	present	A2(24)
15	White Mt.	*L. idas*	24	24	present	A2(15)
16	StrawB Mt.s	*L. idas*	20	20	G	A2(17)
17	Siyeh Cr	*L. idas*	20	20	G	A2(14)
18	Soldier Cr	*L. idas*	20	19	G	A2(12)
19	Tibbs Butte	*L. idas*	20	20	G	A2(17)
20	King's Hill	*L. idas*	18	18	G	A2(12)
21	Garnet Pk	*L. idas*	20	19	G	A2(5)A12(1)B1(2)
22	Shook Mtn	*L. idas*	28	28	present	A2(13)A13(4)A15(1)
23	Wolftone Rd	*L. idas*	4	4	present	A2(3)A13(1)
24	Bunsen Pk	*L. idas*	20	19	G	A2(11)
25	Hayden V	*L. idas*	22	22	G	A2(11)B1(1)
26	Animas RH	*L. idas*	13	13	G	A1(6)A2(2)
27	Red Mt. P	*L. idas*	4	4	G	A1(1)A2(1)
28	Tomboy Rd	*L. idas*	24	24	G	A1(12)
29	Nolan Rd	*L. idas*	8	8	present	
30	Spruce Barley	*L. idas*	20	20	present	A2(1)B1(1)

Table 1 - *Continued from previous page*

#	Locality	Nominal Species	n	# Infected	Data	Wolbchia haplotype
31	Tok	*L. idas*	14	14	present	A2(2)
32	Tolovana Cr	*L. idas*	9	9	present	A2(1)A13(1)A15(1)
33	Soda Mt.	*L. ricei*	20	19	G	A2(12)
34	Rainy Pass	*L. ricei*	20	20	present	A2(12)A17(3)
35	Chinook Pass	*L. ricei*	25	25	present	A2(17)
36	Big Lk	*L. ricei*	20	20	G	A2(10)A3(1)A4(1)B1(5)
37	Cave Lk	*L. ricei*	24	24	G	A1(1)A2(20)
38	Marble Mts.	*L. ricei*	12	7	G	C1(5)G1(1)G2(1)
39	Shovel Cr	*L. ricei*	21	20	G	C1(15)C3(1)
40	Beulah	*L. melissa* - East	10	10	present	A1(1)
41	Brandon	*L. melissa* - East	20	18	C	A1(3)
42	Silver Cr	*L. melissa* - East	6	6	present	
43	Richfield	*L. melissa* - East	6	5	present	A1(2)
44	Victor	*L. melissa* - East	20	20	G	A1(11)
45	Cokeville	*L. melissa* - East	10	10	G	A1(4)
46	Montrose	*L. melissa* - East	20	20	G	A1(9)A16(1)
47	De Beque	*L. melissa* - East	20	19	G	A1(5)
48	Cimarron	*L. melissa* - East	6	6	present	A1(1)A7(1)
49	Goose Lk	*L. melissa* - East	20	20	G	A1(7)
50	Montague	*L. melissa* - East	19	19	G	A1(17)
51	Susanville	*L. melissa* - East	10	10	present	A1(6)
52	Abel Cr	*L. melissa* - East	19	19	C	A1(1)
53	Deeth	*L. melissa* - East	20	20	G	A1(8)
54	Mill Cr	*L. melissa* - East	24	24	present	A1(14)
55	East Cr CG	*L. melissa* - East	25	25	present	A1(8)
56	Lamoille	*L. melissa* - East	20	19	G	A1(10)B1(2)
57	Ophir City	*L. melissa* - East	19	19	G	A1(8)
58	Star Cr	*L. melissa* - East	16	16	G	A1(6)
59	Upper Alkali	*L. melissa* - East	20	19	C	A1(6)A18(2)
60	Surprise V	*L. melissa* - East	20	20	G	A1(13)
61	Cody	*L. melissa* - Rockies	23	22	G	A1(11)A2(1)
62	Lander	*L. melissa* - Rockies	24	23	G	A1(4)
63	Wheatland	*L. melissa* - Rockies	16	16	present	A1(9)A6(1)A19(2)I1(1)
64	Yellow Pine CG	*L. melissa* - Rockies	20	20	G	A1(9)A2(1)
65	Albion Meadow	*L. melissa* - Rockies	46	46	G	A2(40)
66	Lake Davis	*L. melissa* - West	4	4	present	A1(2)
67	Sierravalley	*L. melissa* - West	20	20	present	A1(2)
68	White Lk	*L. melissa* - West	27	27	present	A1(15)A6(4)A11(1)A19(1) B8(1)B44(1)

Table 1 - *Continued from previous page*

#	Locality	Nominal Species	n	# Infected	Data	Wolbchia haplotype
69	Silver Lk	*L. melissa* - West	18	17	G	A1(5)B10(7)
70	Girl Farm	*L. melissa* - West	24	24	present	A1(3)A6(1)A11(2)B5(1) B7(1)B10(1)B18(1)B23(2) B24(1)B25(1)B26(1)B27(1) B28(1)D2(1)E1(1)
71	Verdi Crystal	*L. melissa* - West	73	68	C	A1(14)A6(2)A11(1)A19(1) B5(2)B10(1)B23(1)B29(1) B30(2)B31(1)B32(1)B33(1) B34(1)H1(1)
72	Verdi classic	*L. melissa* - West	26	25	present	A1(2)A19(1)B5(1)B28(1) B35(1)B36(1)B37(1)B38(1) B39(1)B40(1)
73	Verdi tracks	*L. melissa* - West	20	16	present	B10(1)B11(1)B18(1)B22(1) B24(2)B33(1)B41(1)B42(1) B43(1)H2(1)
74	Verdi hwy	*L. melissa* - West	11	11	present	A1(1)A19(2)B23(1)B37(1)
75	Qui	*L. melissa* - West	18	16	present	A6(1)B2(1)B3(1)B4(2) B5(1)B6(1)B7(2)B8(1) B9(1)B10(1)B11(1)
76	Deer Mt Road	*L. melissa* - West	27	23	present	B4(2)B7(1)B12(1)B13(1) B14(1)B15(1)B16(1)B17(1) B18(1)B19(1)B20(1)B21(1) B22(1)
77	Washoe Lk	*L. melissa* - West	20	18	G	A1(2)B10(1)
78	Gardnerville	*L. melissa* - West	18	17	G	B10(6)F1(1)
79	Red Earth	*L. melissa* - West	20	20	G	A1(8)
80	Bishop	*L. melissa* - West	20	20	G	A1(11)
81	Trout Pond	*L. melissa* - West	13	13	C	A1(4)
82	Big Ice	hybrid	18	18	G	A2(11)
83	Blacktail Butte	hybrid	46	45	G	A2(32)
84	Bull Cr	hybrid	46	45	G	A2(27)
85	Dubois	hybrid	41	41	G	A1(1)A2(29)
86	Hunt Mt.	hybrid	30	30	G	A2(24)
87	Periodic Spr	hybrid	20	20	G	A2(28)
88	Pinnacles Butte	hybrid	20	19	G	A2(17)
89	Rendezvous Mt	hybrid	32	32	G	A2(28)
90	Riddle Lk	hybrid	30	28	G	A2(22)
91	Sheffield Cr	hybrid	26	26	G	A2(22)
92	Swift Cr	hybrid	4	3	G	A2(2)

Table 1 - *Continued from previous page*

#	Locality	Nominal Species	n	# Infected	Data	Wolbchia haplotype
93	Buck Mt	hybrid	44	44	G	A2(28)A5(1)
94	Eagle Pk	hybrid	40	40	G	A2(32)A9(1)
95	Steens Mt	hybrid	13	11	G	A2(5)
96	Hinkley	hybrid	26	26	present	A2(21)A13(1)A14(2)
97	Jarbidge	hybrid	42	40	present	A2(30)A11(1)A13(5)A14(2) A15(1)
98	Mt Rose	hybrid	52	8	G	
99	Carson Pass	hybrid	50	32	G	C1(20)
100	Corey Pk	hybrid	8	8	G	
101	Sonora Pass	hybrid	44	33	G	C1(15)
102	Lake Emma	hybrid	33	17	G	C1(8)
103	Sweetwater	hybrid	23	13	G	C1(10)
104	Tioga Crest	hybrid	38	21	G	C1(5)
105	South Fork	hybrid	14	5	G	C1(5)
106	County Line	hybrid	40	35	G	B1(1)B10(6)C1(18)D1(1)
107	Reed Flat	hybrid	9	8	G	C1(4)C2(1)

Table 2: Infection frequencies and strain distributions for *Lycaeides* butterfly species or lineages. Locality numbers correspond with Fig. 1 and Table 1. Dominant strains are the most frequently observed major strains in each lineage (i.e. *wLyc*A, *wLyc*B or *wLyc*C). Minor strains are less frequently observed strains that are often dominant in other lineages and most likely occur in the focal lineage via interspecific transfer. The major subdivisions of strain *wLyc*A, A1 and A2, are treated as strains in this accounting.

Species/Lineage	**Locality**	**n**	**Infection Rate**	Dominant Strain	Minor Strains
L. samuelis	1-6	160	0.51	A1	
L. anna	7-13	115	0.97	C	
L. idas	14-32	333	0.99	A2	A1, B
L. ricei	33-39	142	0.95	A2	A1,C
L. melissa East	40-60	350	0.98	A1	
L. melissa Rockies	61-65	129	0.98	A1	A2
L. melissa West	66-81	359	0.94	A1	B
Jackson Hybrid	82-92	313	0.98	A2	A1
Warner Hybrid	93-95	97	0.97	A2	
Sierra/Whites Hybrid	98-107	311	0.58	C	B

Table 3: Sequence divergence (across 115 variable sites) within and among the three major strains presented as uncorrected percent sequence divergence (p x 100) and (standard deviations).

	***wLyc*A**	***wLyc*B**	***wLyc*C**
***wLyc*A**	3.3% (1.8%)		
***wLyc*B**	37.4% (2.8%)	3.9% (1.7%)	
***wLyc*C**	11.4% (2.2%)	32.1% (2.4%)	3.5% (1.7%)

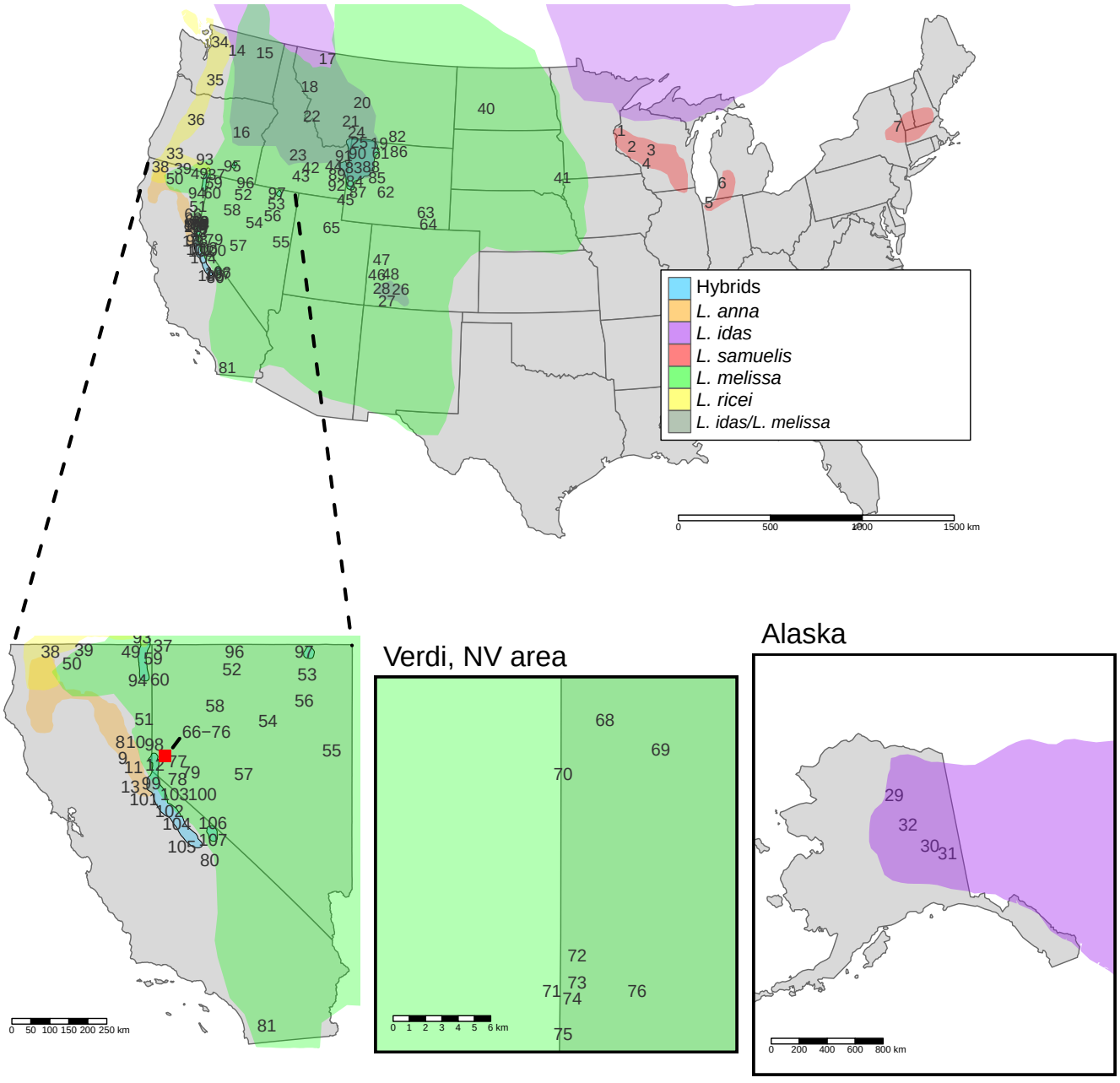

Figure 1: Range maps of the six nominal species of *Lycaeides* in the United States with the 107 sampled locations plotted as site numbers corresponding to Table 1. The dense sampling in the southwestern United States is expanded in the lower left. The red square indicates the Verdi, Nevada sampling area, including sites 66-76, and is also expanded (bottom, middle). Sample locations in Alaska are illustrated in the map on the lower right.

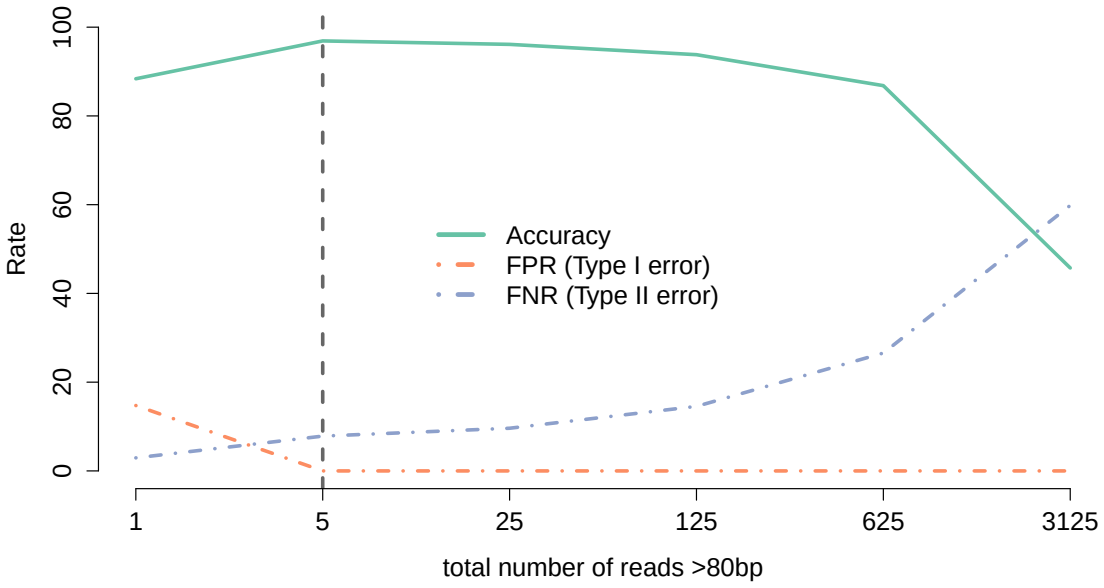

Figure 2: Accuracy and error rates of comparing bioinformatics results to previous PCR-based studies for detecting putative *Wolbachia* infections in the genome for 129 individuals (shown here for a threshold of varying number of reads of length greater than 80 bp). We used a threshold read depth of 5 (five) for classifying an individual as infected, as it had the highest accuracy of 96.9% correspondence with the PCR-based results, while still maintaining a low False Negative Rate (FNR) (classifying an individual as not being infected when the individual is inferred to be infected from PCR-based analysis). False positive rates (FPR) (compared to PCR-based results) were generally low. Note that the X-axis is on a log-scale.

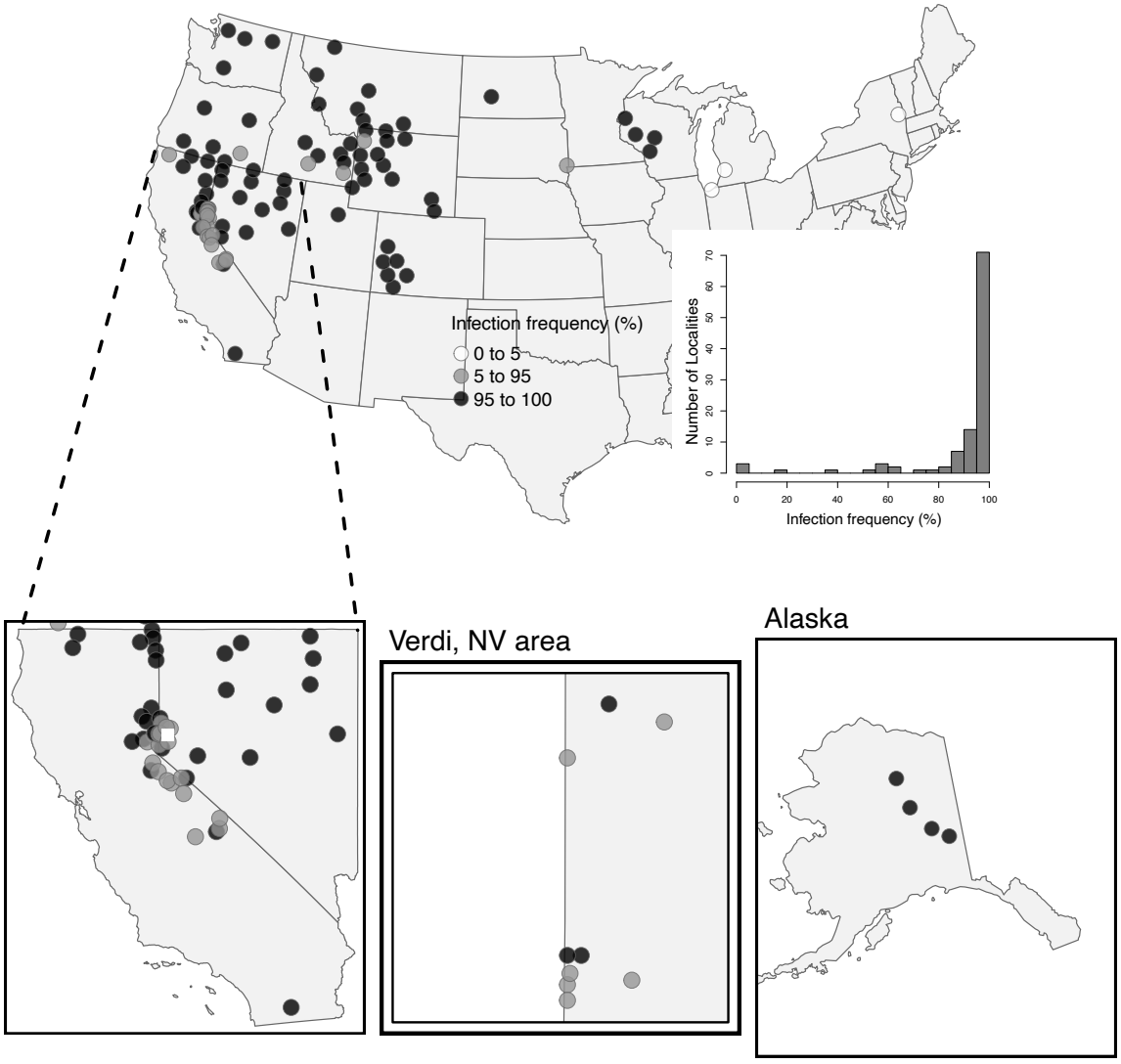

Figure 3: Bubble plots indicating the proportion of infected individuals in a population across the 107 sampled locations. All populations in the western United States are mostly or wholly infected ($> 95\%$), while the *L. samuelis* populations in the east show low to no infection ($< 5\%$). Inset plots zoomed in to regions of interest for visibility. The white square indicates the Verdi, NV sampling area, and is expanded (bottom, middle). Inset plot is a histogram of infection frequencies across 107 sampling localities using a threshold of a minimum of five sequence reads of at least 80bp.

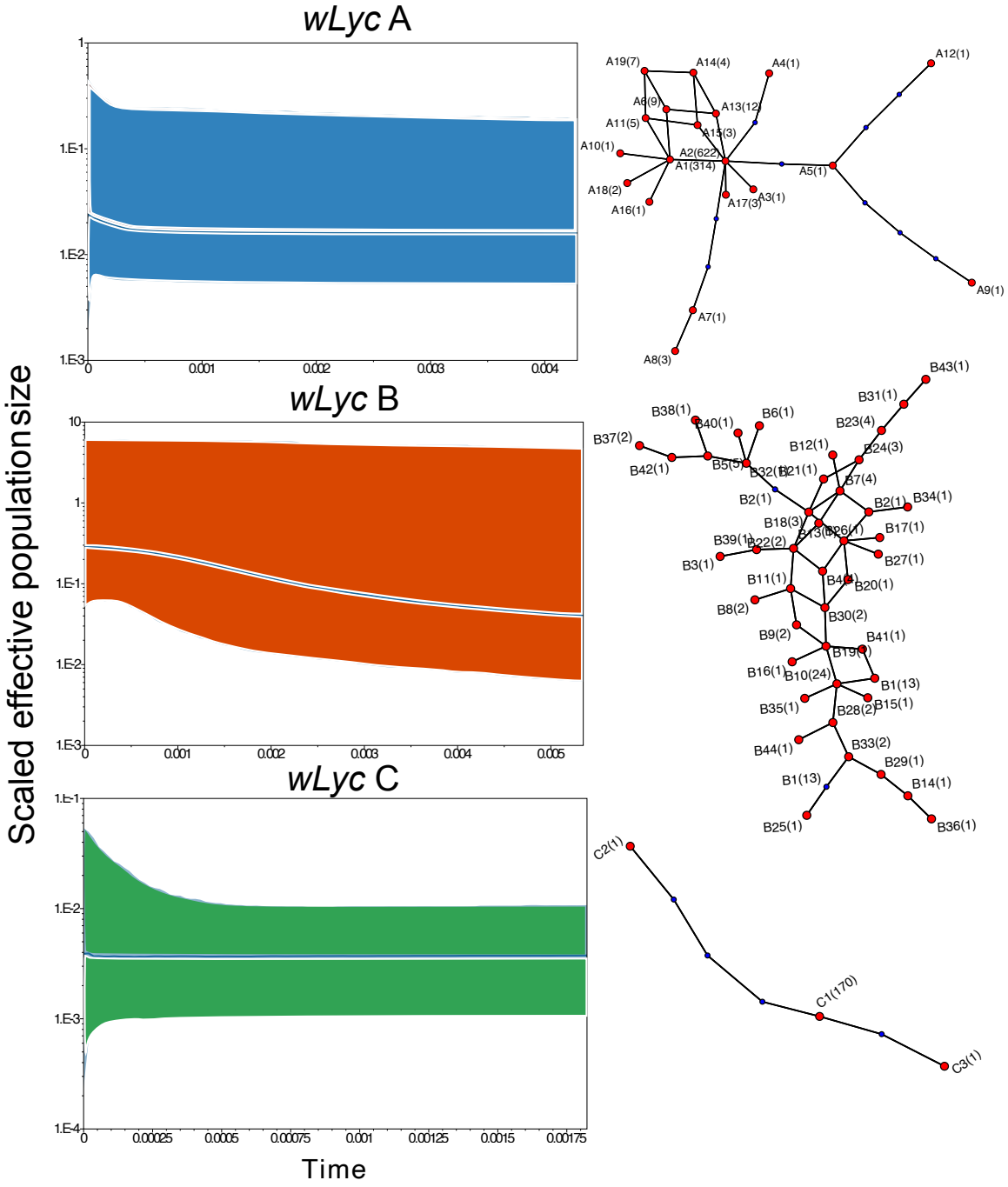

Figure 4: Demographic histories (left) and haplotype networks (right) for each major strain (*wLyc*A, *wLyc*B, *wLyc*C). Population sizes were estimated using `BEAST 2` (Bouckaert *et al.*, 2014). The median mutation-scaled effective population size (dashed line) and 95% credible interval (central posterior density, shaded region) for each strain is presented over time (measured in substitution rate). For simplicity, we assume equal substitution rates across strains to aid interpretation. 95% parsimony networks show observed haplotypes in red and inferred haplotypes in blue with numbers of individuals observed possessing each haplotype in parentheses. Haplotypes are 115bp in length.

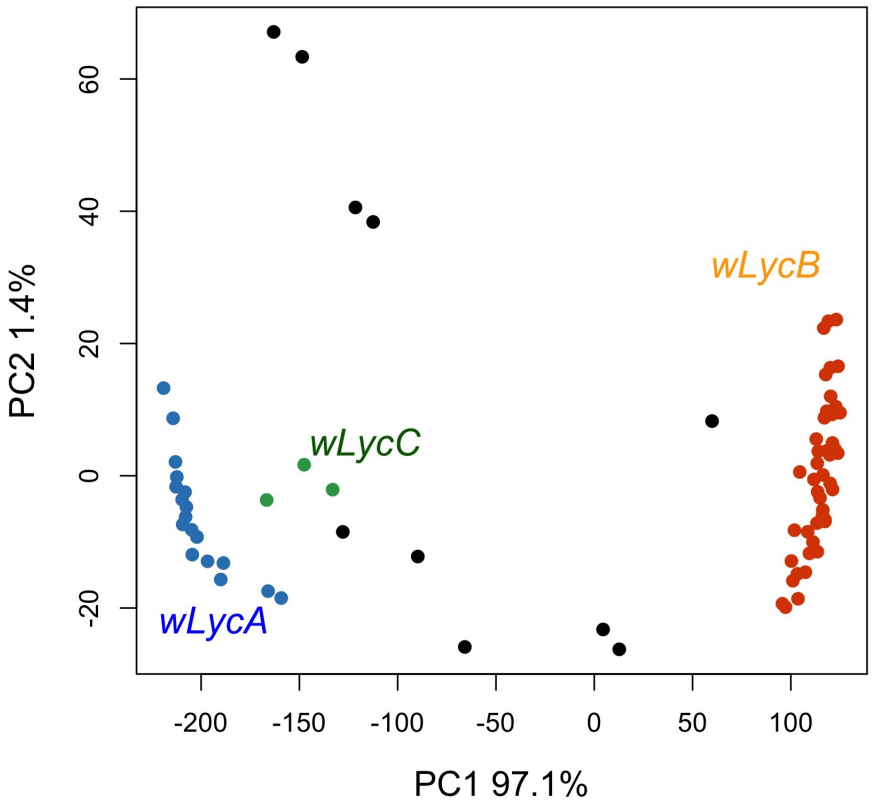

Figure 5: Plot of Principal Coordinates Analysis of *Wolbachia* haplotypes (76 in total) based on uncorrected pairwise distances among haplotypes. Colored dots represent 115bp haplotypes in the three major strains (blue: *wLyc*A, orange: *wLyc*B, green: *wLyc*C). Strain *wLyc*A was found mostly in the *L. melissa*, *L. idas* and *L. samuelis* populations continent-wide. Strain *wLyc*B was mostly found in the *L. melissa* populations in the western Great Basin. Strain wLycC was found exclusively in the *L. anna* populations and in the hybrids between *L. melissa* and *L. anna*. Black dots represent haplotypes found as singletons and not considered part of the three main strains (see Table 1, Supplementary Fig. 3).

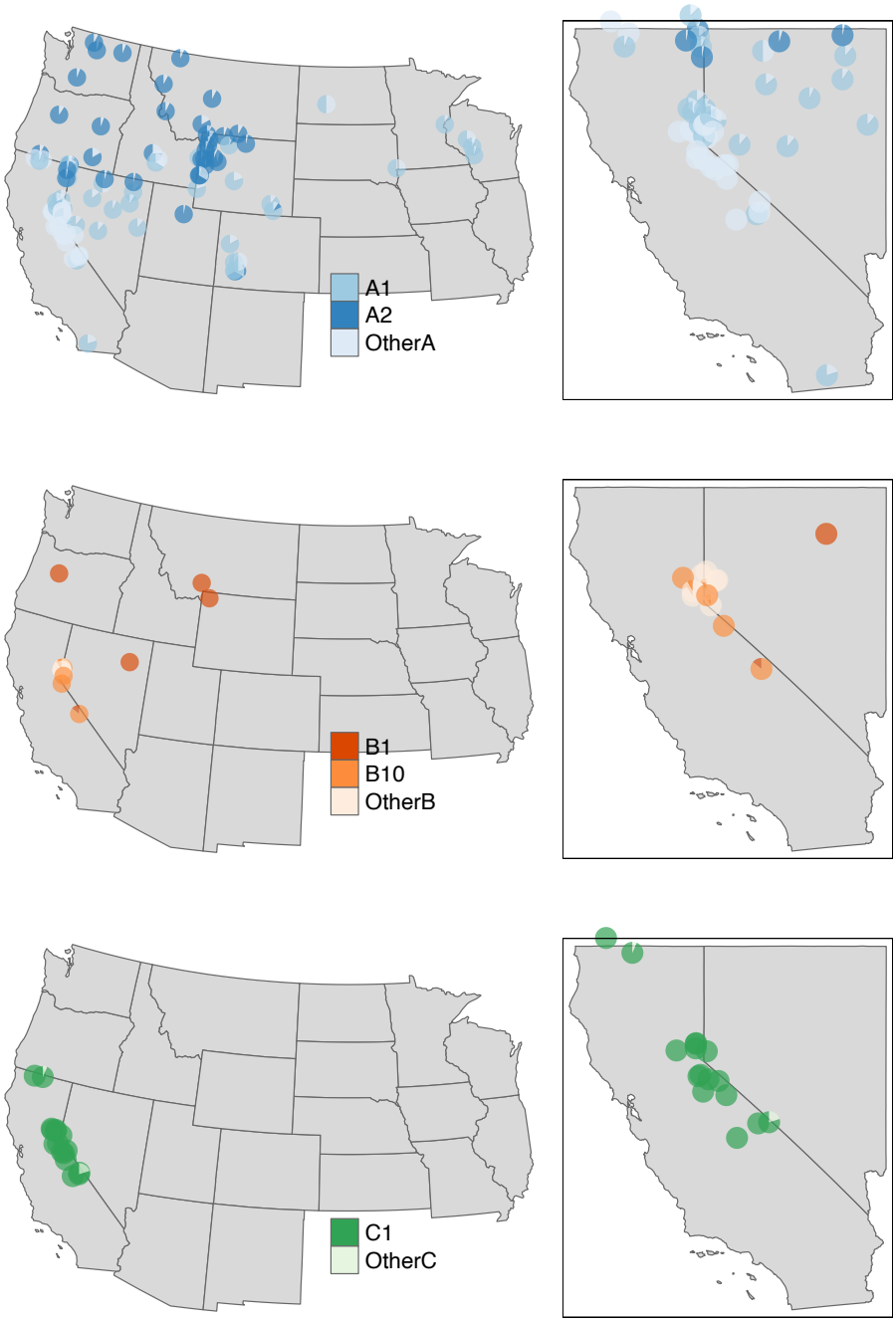

Figure 6: Pie charts showing the distribution of haplotypes from all three strains (row-wise: *wLyc*A, *wLyc*B, *wLyc*C). Haplotypes A1 and A2 are present in 90% of individuals infected with strain *wLyc*A. The label 'OtherA' corresponds to rare haplotypes in *wLyc*A (A3-A19). Haplotypes B5, B9 and B10 make up 78% of all infections in the *wLyc*B strain. Haplotype C1 makes up for 98% of all infections in the *wLyc*C strain, and all other *wLyc*C haplotypes are found in localities that also include haplotype C1. Pies are only shown if a given haplotype is present in the population.